A Summary of Research on and with the STAGES Developmental Model

Tom Murray[1] and Terri O'Fallon[2]

Table of Contents

[1] **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.
tommurray.us@gmail.com

[2] **Terri O'Fallon** is a researcher, teacher, coach, spiritual director and designer of transformative containers. She does ongoing research on the Integral STAGES developmental model which supports a MetAware tier with four later levels of development. Terri is a partner of STAGES International, which creates programs based on the STAGES model. Terri holds a Masters degrees in Special Education and Spiritual direction, and has an Integral PhD in Transformative Learning and Change.
terri@stagesinternational.com

1. Introduction

The STAGES model of adult development is a new framework created by Terri O'Fallon, in consultation with several colleagues over the past decade (O'Fallon, 2010, 2011, 2012, 2013). The model is an extension of the ego-development framework, including the sentence completion test (SCT) mode of assessment, formulated by Jane Loevinger and updated by Susanne Cook-Greuter. It includes elements inspired by Ken Wilber's AQAL model and Sri Aurobindo's model of psychospiritual development (Wilber, 1995; Aurobindo, 1992). The STAGES model diverges from the earlier frameworks in two ways. First, it proposes a small set of underlying factors (parameters or dimensions) that bring about and describe the progression of developmental levels explained in the prior theories. Second, test scorers use an alternative scoring procedure based on these parameters.

Within the past five years, a series of empirical studies have helped launch STAGES from its gestation into a new level of maturity and rigor (a modest step in the usually long process of validating and solidifying any assessment). In this paper, we summarize the results that are reported in detail in several other papers. This paper appears as part of a special issue of the *Integral Review* journal dedicated to the STAGES model. In the "Introduction" article of that issue, we summarize about a dozen empirical research projects, most from within academia, that have used the STAGES assessment or its derivatives as a component of the research project. Given the newness of the model, this count is significant and is an indication of the face validity and usefulness of the model. These STAGES application areas include investigations of organizational change in successful organizations, developmental analysis of women leaders, reflective self-knowledge in healthcare practitioners, psychological resilience in prison inmates, and assessment of the sophistication of climate change understanding.

In this article, we focus on studies that have aimed to validate the STAGES assessment or learn more about its psychometric properties. These studies include:

1. a summary of **prior research** on the sentence completion test (SCT) method;

2. a **replication study** showing that the STAGES model tracks very well with the Maturity Assessment Profile (MAP) method from which it was derived, up to 4.5/Strategist (Replication was not attempted above this level for the reasons described).;

3. studies showing that the STAGES assessment remains robust for **variations in sentence stems and test lengths**;

4. l**ongitudinal** data analysis that shows, among other things, that the MetAware (Tier 3) stages exhibit appropriate developmental progression;

5. additional analysis of **late stage patterns** that shows that confirms that those scoring in the Metaware tier have a substantial number of test items scored at the complex 4.5/Strategist level;

6. additional inter-rater analysis, including more recently trained scorers and per-item inter-rater reliability (IRR) statistics;

7. a research study that uses **item response theory** and Rasch analysis to investigate many psychometric properties of the STAGES assessment and that elaborates on a number of **problems with the** ogive item-aggregation method while also suggesting an alternative;

8. a "**Face Validity**" section describing various uses of the STAGES model, including its use in 10 dissertation or thesis projects.

9. a description of preliminary analysis of how specific **words/concepts** appear as **frequency distributions** across the stage spectrum; and

10. description of a study on **how children respond** to the sentence completion test.

In several of our research papers, we make a point of noting that positive results on the STAGES assessment generally transfer to positive results for the Loevinger lineage of SCTs, including Cook-Greuter's and Torbert's MAP, Global Leadership Profile (GLP), and Leadership Development Profile (LDP) SCTs. Our articles also discuss properties that differentiate STAGES from the other assessments in this lineage. However, many of the conclusions we draw (e.g., regarding the validity of much shorter versions of the SCT and the validity of using alternative sentence stems) can be applied by other SCT developers.

2. The STAGES Model

The SCT for adult development, created by Jane Loevinger and later extended by Susanne Cook-Greuter, William Torbert, and Terri O'Fallon, 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 & Wesler, 1970, p. 3).[3] Ego development has been described using a variety of concepts including leadership maturity, perspective-taking complexity, sophistication of world-view consciousness, and "wisdom skills." It has substantial overlap with the construct of meaning-making maturity described in Kegan's construct developmental theory (Kegan, 1994). Kegan describes this construct 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, others, and the world – intrapersonal, interpersonal, and cognitive in Kegan's terms and "I/we/it" in terms of Wilber's Integral Theory (1995).

The SCT is a "projective" test in which subjects complete sentence starters, responding freely without a need to produce a "correct" or superior answer – which the theory claims affords the test analyst a deeper look into tacit or unconscious psychological traits. The standard SCT has 36

[3] 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."

independently rated sentence starters, or "stems" (e.g., "Raising a family…" and "When people are helpless..."), which the test taker completes (e.g., "…is a joy" and "…they get taken advantage of"). Sentence completions vary from a few words (or even one word) to full paragraphs (and, rarely, multiple paragraphs). Rather than taking a simple sum or mean of the item scores, the total protocol rating (TPR) score uses a more complex cutoff method called the "ogive method" for reasons we describe later. The ogive method converts a complex, continuous score into one of a set of discrete categories (i.e., levels or stages): 8 for Loevinger's model, 9 for Cook-Greuter's and Torbert's models, and 12 for O'Fallon's model.

Regarding the theoretical model, Cook-Greuter extended Loevinger's system by differentiating two stages (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, differentiating four stages (5.0, 5.5, 6.0, and 6.5) within the same late-stage territory. O'Fallon also differentiates the Conformist (or Diplomat) level into two levels, basing this and her other modifications on theoretical principles. Thus, the STAGES model defines 12 levels, MAP/GLP/LDP defines 9, and the Washington University Sentence Completion Test (WUSCT) defines 8.

In this issue, other papers, including O'Fallon et al. (2020) and Barta's "Psychological Application of the STAGES Model," describe the STAGES model in detail. Figure 1 illustrates the three questions that define the model.

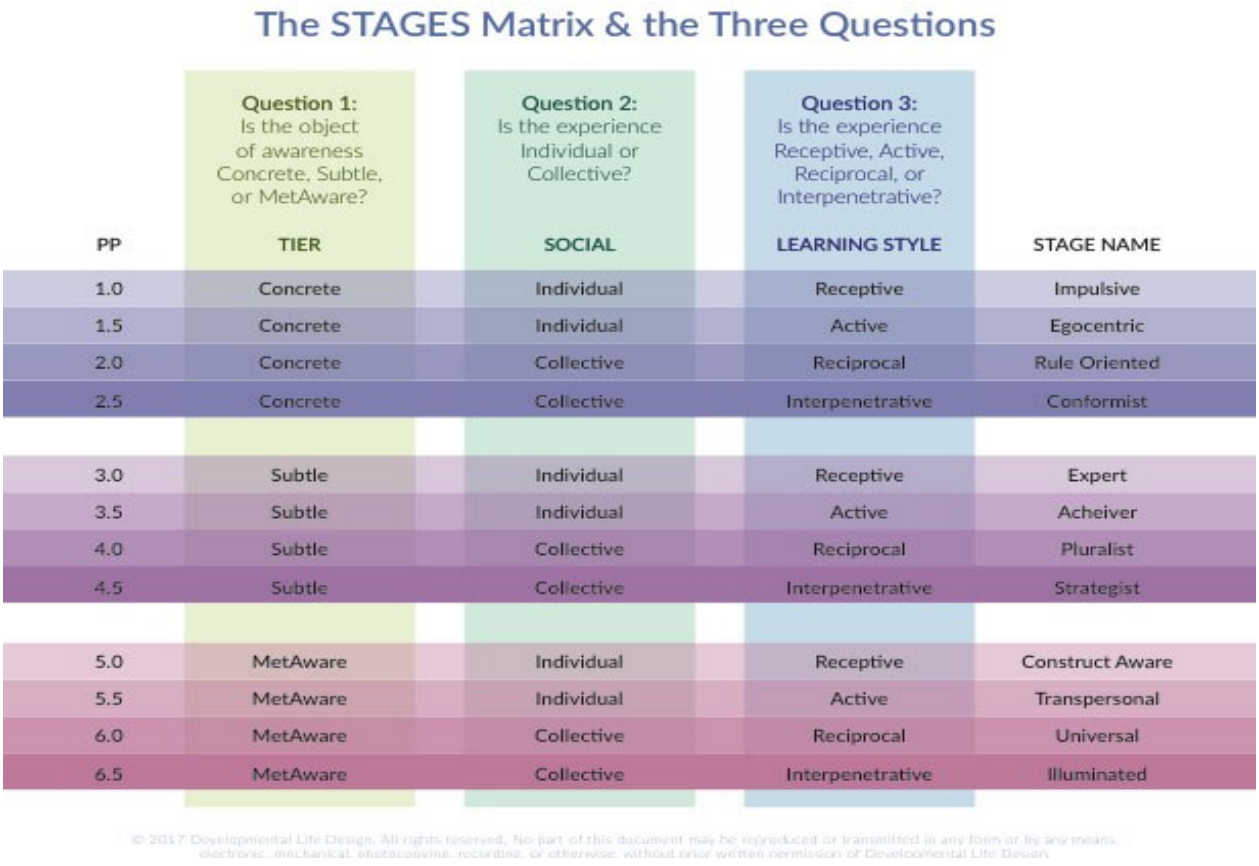

Figure 1. The STAGES model.

3. Validity of the Sentence Completion Test

[This paper contains excerpts from "Sentence completion assessments for ego development, meaning-making, and wisdom maturity, including STAGES," by Murray (2017).]

Murray (2017) contains a lengthy summary of validity and reliability studies of the SCT, supporting Westenberg et al.'s conclusion that "findings of over 350 empirical studies generally support critical assumptions underlying the ego development construct" (p. 485). Below we provide a summary of the main outcomes of this history of research on the SCT for inter-rater reliability, internal consistency, test-retest reliability, face validity, construct validity, incremental validity, clinical utility, external validity, and predictive validity (included in text copied from Murray (2020) in this issue). In later sections, we discuss the analysis of the STAGES model.

Of the more than 400 studies of the SCT, the majority were conducted with Loevinger's WUSCT, but studies of the other variations inevitably replicate the results of the WUSCT's general validity characteristics. Torbert (2014) summarizes a number of studies on the MAP, GLP, and LDP instruments that postdate the WUSCT studies.

Loevinger's theory of ego development was intricately linked to her assessment instrument, the WUSCT. (Hy & Loevinger, 1989; Loevinger & Wessler, 1970). The literature on the SCT includes over 40 years of meta-analyses and critical overviews, which substantially support its validity and usefulness (see Cohn & Westenberg, 2004; Manners & Durkin, 2001; Holt, 1980; Novy & Francis, 1992; Jespersen et al. 2013; Westenberg et al., 2004b; Forman, 2010). According to an overview by Westenberg et al. (2004a), the SCT has robust psychometric properties, having "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 of studies have followed since 2004 (Torbert, 2014).

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 reasons given above, we claim that these results apply to STAGES as well. Results of the meta-analyses include the following:

- "Psychometric studies of the WUSCT…invariably report high levels of inter-rater reliability" (Westenberg et al., 2004a, p. 603; see Torbert & Livne-Tarandach, 2009 for strong results on the MAP, GLP, and LDP; and see 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; 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 socioeconomic status (SES) measurements (Browning, 1987; Cohn & Westenberg, 2004).

- The SCT has proved applicable in different countries, **cultures, and languages** (see Carlson & Westenberg, 1998).

- Though 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 she 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).

4. STAGES vs. MAP Replication Study

*[This section contains excerpts from The Validation of a New Scoring Method for Assessing Ego Development Based on Three Dimensions of Language, by O'Fallon et al., 2020: [*https://doi.org/10.1016/j.heliyon.2020.e03472*]*

The first major study of the validity of the STAGES assessment method was a "replication study" comparing the STAGES model to the MAP model (from Cook-Greuter), from which STAGES was derived. The main purpose of the study was to show that SCTs scored by both systems would have a high reproducibility (i.e., to show that STAGES measures essentially the same general construct) while using an assessment methodology that is, according to its developers, more flexible, efficient, and explanatory.

Because the two systems have relatively different definitions of levels above 4.5/Strategist, the replicability study was conducted for stages up to and including 4.5. We called this the Tier 1-2 data set, and the data for levels higher than 4.5 was called the Tier 3 data set (Metaware tier in the STAGES model). Note that according to very rough estimates of the general population, Tiers 1

and 2 combined represent approximately 98% of all adults and 92% of all professionals (Cook-Greuter, 2004, p. 279; Torbert, 2009).[4]

Because of the expected divergence in level definitions in Tier 3, for that tier, we conducted an IRR study as an indicator of test consistency (see the longitudinal analysis below for evidence of Tier 3 validity). The study included an IRR analysis for the full set of inventories, combining Tiers 1 and 2 with Tier 3.

Method. From a set of approximately 750 inventories, most of which were scored previously using the Cook-Greuter/Loevinger (CG/L) method, 142 were selected for this study using stratified methods described in the main paper. The selection criteria considered that the goal was to have sufficient representation in each of the 32 stratified sampling categories and to have approximately 12 inventories at each STAGES level.

For this study, each of these inventories was scored by three STAGES scorers using random assignment of inventories to four certified scorers (i.e., there were four scorers with inventories assigned such that each inventory was scored thrice).[5] The MAP scores were taken as the gold standard against which STAGES scoring would be measured (the IRR of the MAP scoring system had been previously demonstrated). Having four independent raters for the STAGES system comprises a particularly rigorous method of testing. We compared the STAGES model and MAP compared both per rater and for all raters collectively. Because one of the scorers ("scorer #1") was in the unique position of having learned both the MAP and STAGES scoring methods, a separate analysis was performed, aggregating the three other scorers as well as looking at all four scorers together.

The level of agreement for Tier 1-2 data was quantified by the weighted Cohen's Kappa statistic (Cohen, 1968). Using the Kappa statistic, we compared the STAGES scoring for each of the three scorers separately with the single CG/L score. Furthermore, we calculated the mean Kappa values across all scorers.

Results. For the Tier1-2 study (up to 4.5), there was "substantial" or "excellent" agreement[6] for all methods of evaluating STAGES vs. MAP scoring (i.e., for each scorer, for all combined, and for all except scorer #1).

For the Tier-3 study, including levels 5.0, 5.5, 6.0, and 6.5, the overall agreement among the raters was "substantial" and the agreement was "moderate" for the analysis with the more experienced scorer #1 excluded.

[4] However, for the clients served and scored by Pacific Integral, the percent scored in Tier 3 is higher, approximately 25%.

[5] To date, approximately10 individuals have been certified to score using the STAGES model, and more are under supervision for certification. The first cohort of four trained scorers, now referred to as master scorers, participated in the validity study described in this research.

[6] Using a widely referenced set of labels, Kappa values can be interpreted as follows: $\kappa < 0.0$, no agreement; $\kappa = 0.0–0.20$, slight agreement; $\kappa = 0.21–0.40$, fair agreement; $\kappa = 0.41–0.60$, moderate agreement; $\kappa = 0.61–0.80$, substantial agreement; and $\kappa = 0.81–1.00$, excellent agreement (Landis, 1977).

For the all-tier (Tiers 1, 2, and 3) study, the IRR among scorers across all 12 stages was strong both for all scorers and where the more experienced scorer #1 was excluded.

The results above summarize the overall agreement between the two models. At a more detailed level, we might ask how this agreement looks at each developmental level. Appendix 4 of the main article includes a detailed look at the per-level agreements, as shown in Table 1. That paper explains how the Tier 1-2 scores of all raters were combined to produce the STAGES scores given in the table.

Table 1. Agreement details for STAGES vs. Cook-Greuter/Loevinger systems.

(CG/L)*	1.0 (2)	1.5 (2/3)	2.0-2.5 (3)	3.0 (3/4)	3.5 (4)	4.0 (4/5)	4.5 (5)	**Total**	**Agrmnt**	**Agree +/- 1**
STAGES										
1.0	**4**	0	0	0	0	0	0	4	100.0%	100.0%
1.5	5	**5**	3	2	0	0	0	15	33.3%	86.7%
2.0-2.5	3	5	**35**	9	1	0	0	53	66.0%	92.5%
3.0	0	2	9	**21**	1	0	0	33	63.6%	93.9%
3.5	0	0	1	4	**28**	7	3	43	65.1%	90.7%
4.0	0	0	0	0	3	**19**	6	28	67.9%	100.0%
4.5	0	0	2	0	3	9	**23**	37	62.2%	86.5%
5.0	0	0	0	0	0	1	4	5		
Total	12	12	50	36	36	36	36	218		
Agrmnt	33.3%	41.7%	70.0%	58.3%	77.8 %	52.8%	63.9%			
Agree +/- 1	75.0%	83.3%	94.0%	94.4%	88.9 %	97.2%	91.7%			

(*The CG/L stage is noted in parentheses).

Table 1 shows the number of STAGES scores assigned to each CG/L score for each possible STAGES score. For example, the first item in the table with a count of "4" indicates that, totaled over all four scorers, there were four instances of a CG/L score "2" with a STAGES score of "1.0." The table shows percent agreement for each row and column and includes an agreement ("+/- 1") within a one-level window. The total number of inventory scores was 218 across the four scorers. As the CG/L system does not differentiate 2.0 and 2.5, these were combined. The agreement numbers can be interpreted in relation to the fact that certified scorers in both systems exhibited a margin of error – usually at least 85% agreement between accuracy and a master scorer.

From these detailed results, we concluded the following:

- Similar to the analysis of the Tier 1-2 study, the overall weighted Kappa score for this matrix was 76%, well into the "substantial agreement" level.

- The number of exact matches (values along the diagonal) varied per level. For levels above 1.5, there was a substantial number of exact matches, but for levels 1.0 and 1.5, the number of matches was unacceptably low relative to the number of inventories rated at this level by CG/L). However, in real applications, less than 1–2% of adult individuals were expected to score at these levels, making this mismatch insignificant.

- The "within one level" metrics were all high, ranging from 86.5% to 100% for levels above 1.5.

The STAGES model seemed to bias the following levels marginally higher compared to CG/L – 2.0–2.5, 3.5, 4.0 – while it seemed to bias levels 3.0 and 4.5 slightly *lower* than CG/L. Colleagues have expressed anecdotal concerns that the STAGES model would bias the highest developmental levels higher than the CG/L model, thus giving participants a false sense of higher development. Our data suggest a small degree of such an effect in assigning CG/L Pluralist (4.0) scores into Strategist (4.5), but it shows a reversal of this trend for the highest level where the bias is for STAGES to rate CG/L Strategist individuals as Pluralist, or lower. Overall, there does not appear to be a significant shift, higher or lower, in STAGES vs. CG/L scoring. As mentioned in the description of the Tier 1-2 study, at levels above 4.5, the available data did not allow for a robust comparison, and this question remains for future research.

Overall the study supported the hypothesis that STAGES replicates MAP scoring, and is thus comparable to prior measurement methods in the Loevinger tradition, up through 4.5/Strategist. Later research described below adds substantially to these results, in terms of inter-rater reliability (section 8), and the validity of measurements in the Metaware tier (sections 6, 7).

5. Modifications to Test Length and Sentence Stems: Internal Consistency

[This section contains excerpts from two papers in this journal issue: "The STAGES Specialty Inventories: Robustness to Variations in Sentence Stems Integral Review," by O'Fallon & Murray and "Investigating the Validity of the Ogive Aggregation Method, Including the use of Rasch Analysis, for the Sentence Completion Test and the STAGES Model," by Murray.]

Using the Cronbach's alpha statistic, prior research on the reliability of the WUSCT and its derivatives (MAP, GLP, and LDP) has consistently shown good to excellent results for internal consistency. Prior studies have evaluated specific inventories, and we have extended this research with general conclusions regarding the robustness of the SCT to changes in sentence stems.

The STAGES model allows for efficient introduction and experimentation with variations of the SCT, including new sentence starters and length variations. O'Fallon and colleagues have developed and evaluated "specialty protocols" in these domains: leadership, love, education,

climate change, and a children's version of the SCT. Most of the specialty protocols replace six of the general protocol stems with theme-specific stems. In this issues' "Specialty Inventories" paper, we show the results of the internal-consistency studies of (a) the STAGES general protocol, (b) six specialty protocols, and (c) a shorter 16-item inventory.

Next we describe reliability analysis of the six new ("specialty") inventories: leadership, love, education, psychological reflection, climate change, and a children's inventory For the children's inventory, 53 children, age range 4-13 enrolled in a progressive elementary school in Brisbane Australia, gave verbal answers to in-person prompts; recordings of their answers were transcribed. The data is shown in Table 2.

Table 2. Cronbach's alpha internal consistency values for specialty protocols.

	Cronbach's alpha					
Domain	**Leadership**	**Love**	**Education**	**Climate Change**	**Psych**	**Children**
N in study	32	20	20	32	17	53
New stems (of 36)	6	6	6	6	5	11
α All 36 stems	0.97	0.95	0.95	0.96	0.92	0.88
α General prot. stems only	0.96	0.94	0.94	0.95	0.90	0.85
α New stems only	0.85	0.82	0.83	0.82	0.80	0.63

Conclusions: For the five specialty inventories excluding Children's:

- the overall internal consistency was excellent (α 0.95 to 0.97).
- the internal consistency for the new stems as a group was good (α 0.80 to 0.85).
- For the Children's inventory:
- the overall internal consistency was good (α 0.88).
- the internal consistency for the new stems as a group was questionable (α 0.63).

Overall, these results not only show evidence about the reliability of the particular specialty inventories, but it is also show evidence that the SCT is quite robust to the addition of new stems, assuming they are well-written. It also gives evidence that even a short test containing only six specialty stems would be a psychometrically reliable instrument.

The Children's protocol is a special case. First, there were almost twice as many new stems as the other specialty inventories. Second, the assessment was done face to face and verbally. Third, the pre-existing (general protocol) stems by themselves show a much lower alpha (though still "good") vs. the other protocols, indicating that there was probably much more variation in the children's responses than in adult's (see O'Fallon's paper on the Chidren's protocol in this issue). Thus, though the alpha of the new stems by themselves was unacceptably low, it is not clear whether this was because of the nature of the stems or the nature of the subject population.

Additional analysis (the Cronbach-Mesbah Curve in "OGIVE and Rasch" analysis paper in this issue) also shows that the STAGES assessment (and the SCT in general) is psychometrically valid for much shorter inventories, as low as 10 and even 5 items for certain applications.

In addition, we are currently working with domain experts to design additional specialty protocols in these areas: relationships, religious beliefs, spirituality, money, hope, dementia, ethics, and parenting.

6. Longitudinal Analysis

[This section contains previously unpublished data and results summarized in "The Validation of a New Scoring Method for Assessing Ego Development Based on Three Dimensions of Language," by O'Fallon et al., 2020: https://doi.org/10.1016/j.heliyon.2020.e03472]

In the current database of STAGES scores, a more recent data set than that used for the replication study above, 115 individuals have taken the assessment more than once. Evidence that each subsequent test is highly likely to yield a score equivalent to or higher than the previous score (i.e., monotonic growth) is considered strong evidence of a "developmental" construct. Below we summarize our longitudinal data analysis for the developmental spectrum while also focusing on the MetAware tier.

Of the 1,245 surveys in the database, 143 were retests representing 115 clients, 88 having taken one retest, 20 having taken two retests, 5 having taken three retests, and 3 having taken four or five retests.[7] The average time difference between retests was 2.1 years. In this analysis, we ignored the time differences between tests. (For future analysis, we will also factor in retest gap time using multilevel modeling).

Table 3 includes a summary of the data showing the prior and next score of each retest. Table 4 shows the same data in terms of the amount of change. Each of the 143 retests was considered as an independent event; 38% remained the same, 50% increased, and 11% decreased. Thus, 89% either increased or remained the same. The 11% that decreased could be explained by a combination of factors and "noise" including rater error, test-retest variability (i.e., tests taken on the same day had some percentage chance of differing), or actual "regressions" due to major life challenges resulting in cognitive or emotional stressors. Gains could be attributed to test "practice effects," but the 2-year average gap between retests makes that highly unlikely. These results constitute substantial evidence corroborating prior research that shows the ego development construct (with WUSCT) is developmental in nature – now shown for the STAGES assessment as well.

[7] The few retests that were less than three months apart were excluded.

Table 3. Longitudinal: prior vs. next scores.

	Next Score							
Prior Score	3.5	4	4.5	5	5.5	6	6.5	Grand Total
3.5	3	6	2	1				12
4	2	6	8	3				19
4.5	1	3	21	27	6			58
5			4	15	12	1		32
5.5	2		1	2	9	5		19
6					1	1	1	3
Grand Total	**8**	**15**	**36**	**48**	**28**	**7**	**1**	**143**

Table 4. Longitudinal: prior scores vs. amount of change to next score.

	Change to Next Score							
Prior Score	-2	-1	-0.5	0	0.5	1	1.5	Grand Total
3.5				3	6	2	1	12
4			2	6	8	3		19
4.5		1	3	21	27	6		58
5			4	15	12	1		32
5.5	2	1	2	9	5			19
6			1	1	1			3
Grand Total	**2**	**2**	**12**	**55**	**59**	**12**	**1**	**143**

Many of the test subjects entered a program aimed at personal and professional growth called "Generating Transformational Change," or GTC. This program included developmental models as part of the curriculum. The test subjects may have learned vocabulary that led to higher verbal-textual SCT scores without advancing their deeper "enactive or embodied" development. (The research, assessment tools, and adequate theory do not yet exist for allowing one to separate the verbal and non-verbal components of developmental change. If we focus solely on the 47 retests from non-GTC subjects, we see that only 17% of the retests led to a decrease in scores, again confirming the developmental nature of the ego development construct. The GTC cohort shows the most improvement overall, with only 8% of the retests resulting in decreasing scores).

Finally, we can focus our longitudinal analysis on the third tier, MetAware, which was excluded from our replication study with the L/CG data for reasons explained above. Table 5 shows compelling evidence for the developmental sequencing of O'Fallon's newly defined highest stages. Of the 84 retests for which the score was in the MetAware tier, 67% increased,

30% stayed the same, and only 4% decreased – i.e., 96% increased or stayed the same. These scores are an even stronger indicator of monotonic sequencing compared to the scores of all three tiers combined. The data is a strong indicator that, developmentally, each MetAware stage, sits after its STAGES precursor and before its STAGES successor.

Table 5. Longitudinal data when the new score is in the MetAware tier.

	Score Change					
Prior Score	-0.5	0	0.5	1	1.5	Grand Total
3.5					1	1
4				3		3
4.5			27	6		33
5		15	12	1		28
5.5	2	9	5			16
6	1	1	1			3
Grand Total	**3**	**25**	**45**	**10**	**1**	**84**

(More evidence related to the Metaware tier is in the next section).

7. Analysis of Late Stage Patterns

[This section comes from the paper "Deconstructing Developmental Constructs: A conversation with Thomas Jordan and Tom Murray" in this issue.]

We have been engaging in conversations with Thomas Jordan (see the article this issue), who provided a critical inquiry into the STAGES model. After sharing our longitudinal data (above) with Jordan, he wanted a deeper look into the data to ask the question of whether the Metaware tier (as STAGES defines it) was necessarily built ("stacked") on top of the subtle tier (and did not seem to skip past it). To put it another way, he wanted to know whether those scoring in the Metaware tier also showed that they have the complexity associated with 4.5/Strategist, and did not present as a group displaying a lot of esoteric or spiritual language with little evidence of the rational, analytic, or self-reflective rigor needed to move into Strategist. To answer this, we ran an additional analysis to show the overall distribution of item scores for those who scored in the Metaware Tier. The figure below shows, for protocols with center of gravity values (ogives) of 3.0 to 6.5, what percent of the 36 item scores are at each level.

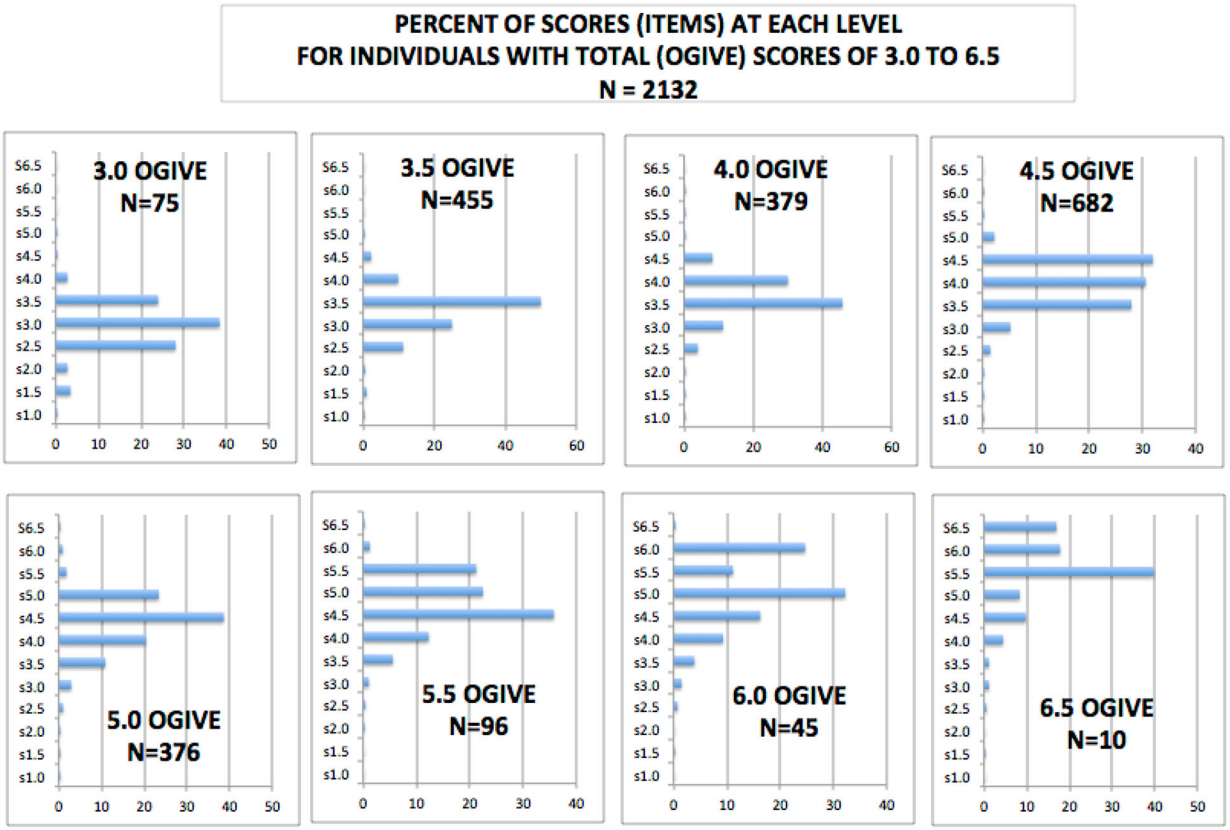

Figure 2. Percent of scores at each level for Metaware surveys.

Discussion of results:

- You can see that for 5.0, 5.5, and 6.0 there are many scores in 4.5 (Strategist). (Recall that the ogive cutoff method prioritizes the higher scores, so the total score is not the average, and is often not the mode, i.e. top of the bell curve).

- However, the scores do tend to follow a normal distribution (bell shape) around the average, so at 6.0 and 6.5 there are naturally less and less in the Subtle tier.

- The 6.5 graph by itself might seem to confirm a suspicion that the metaware tier does not have many subtle tier scores, but (1) there are not as many scores there so its a weaker data point, and (2) people tend to write close to their center of gravity or average, even if the have the capacity to write lower.

- Recall that with each tier the type of object changes, so the complexity *about that object* can start again at low and build up from there – so there may actually be a kind of reduction in horizontal complexity starting at 5.0 (as in Fischer's model, in which, for example, "single principles" follows "systems of abstractions").

- We are finding that in general the scoring system puts more in the X.5 (active) vs the X.0 (passive) scores. We think this is because people in passive phases have a harder time finding he words, and drop down to the prior level for content.

- Compared to a normal distribution, 4.5 scores have a lot of 3.5 scores, 5.5 surveys have a lot of 4.5 scores, 6.0 surveys have a lot of 5.0 scores; and 6.5 surveys have a lot of 5.5 scores – this many indicate that for higher levels, a person with a active (X.5) center of gravity uses more active language in general (and vice versa for passive (X.0)).

In sum, scores in the metaware tier (at least 5.0, 5.5, and 6.0) have substantial scores in the subtle tier, and in particular scores at 4.5/Strategist that tend to have high complexity.

8. Additional Inter-Rater Studies

[This section contains data and results not published previously, but summarized in "The Validation of a New Scoring Method for Assessing Ego Development Based on Three Dimensions of Language," by O'Fallon et al., 2020: https://doi.org/10.1016/j.heliyon.2020.e03472.]

The replication study described above included an IRR statistic for all levels, including the MetAware tier. That study was conducted using the first cohort of four trained STAGES scorers. As O'Fallon has continued to train scorers over the succeeding 4 years, we have additional data on the IRR strength of the instrument. The original study was a comparison at the survey level only because data at the per-item level was not available for the MAP scores. The more recent data includes per-item IRR statistics as well.

STAGES scorers are "certified" after completing a training program, which takes about one year and includes the supervised scoring of approximately 100 inventories. The scorers then practice their skills until they achieve greater than or equal to 85% correct scoring, as compared with a master scorer, for 10 consecutive inventories.[8] This is for agreement at the inventory level.[9] To obtain an additional indication of the inter-rater reliability of the scoring method, we can assess the item-level, or stem-level, agreement. We have data for the five most recently certified scorers trained over the last three years. Over an aggregate of 36 scores, the final 10 pre-certification scores showed that these scorers had a survey-level accuracy considerably higher than the 85% minimum requirement. Of the 50 surveys (10 each for five scorers), only one did

[8] All the statistics are comparing the scorer to the expert. We cannot compare scorers to one another (i.e., obtain a multi-scorer inter-rating) because each scorer's list of final 10 surveys is unique; they did not score the same surveys.

[9] Once a scorer is certified, STAGES International has several means of maintaining quality control with its scorers. First, there is an active email discussion list in which scorers post completions that are uncertain about how to allow the community to reflect on them and learn together. Second, the scoring procedure (scoring manual) includes additional secondary features of the text that are used to cross-check final scores, including a rubric of common concepts and terms used at each level. Finally, inter-ratings are periodically preformed to check scorer accuracy vs. a master scorer.

not have perfect accuracy based on the previously established gold standard. (because one incorrect result was one level off).. Thus, the overall accuracy at the survey level was 98%.[10]

Table 6. Rater accuracy aggregated over 10 scorings (of the 36-item test) for each rater.

Scorer	Average Acc	Min Acc
1	0.88	0.72
4	0.93	0.83
3	0.91	0.86
2	0.94	0.89
5	0.97	0.92
Average	0.93	

At the item level, agreement was also excellent (see Table 6). The average accuracy was 93%. When taking the average accuracy over their 10 scores, the highest was 97% and the lowest was 88%. In the set of 50 survey scores, the lowest average-over-items accuracy was 72%, and the highest was 100% (the majority of errors were off by one level). Four of the five scorers had at least 2 of the 10 surveys at 100% stem-level accuracy. This is strong evidence for the reliability of the scoring method, which, comparing these numbers to the main study in this paper, has improved in recent years, likely due to improvements in the training program.

9. Investigating the Validity of the Ogive Aggregation Method, Including Rasch Analysis

[This section contains excerpts from "Investigating the Validity of the Ogive Aggregation Method, Including the use of Rasch Analysis, for the Sentence Completion Test and the STAGES model," by Tom Murray, which is part of this issue – specifically, this section includes the executive summary from that paper.]

In this issue, Murray reports on an extended project to assess the psychometric properties of the STAGES SCT using data from 740 scored surveys. The goals of the research are to (1) 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 (2) to further investigate suspected problems with and propose alternatives for the ogive cutoff method for aggregating item scores in the SCT.

We focus on methods used to aggregate the 36 item scores into the total protocol rating (TPR), chief among these being the ogive cutoff method, and secondarily the less-frequently used total

[10] We use percent accuracy rather than Cohens Kappa here. The Kappa statistic is not well suited for high accuracies—see the "Kappa paradox" (Feinstein et al., 1990).

weighted score (TWS) method. We argue in detail that the ogive method has a number of drawbacks, including:

1. a lack of empirical support for discontinuous levels,
2. no population studies for Bayesian estimates,
3. difficulties estimating extreme values,
4. lower cutoffs that confuse development with "shadow" evidence,
5. sensitivity to error,
6. TWS misalignment,
7. additional issues with the TWS multipliers, and
8. errors in the ogive formula and assumptions.

The paper then describes the method and results of our statistical analysis of the 740 STAGES protocols using general descriptive statistics, IRT, and Rasch Analysis (a subset of IRT). Below is a summary of the findings. These findings hold for levels 2.0 to 6.0. For the rare extremes of 1.0, 1.5, and 6.5 surveys, there are insufficient data points for drawing confident conclusions.

1. **Within-test item normality**. A skew and kurtosis analysis show that scores within each survey are, on average, normally distributed. That is, the supposition that the scores "bell curve" within a survey would contain more items below the average score than above it is not supported.

2. **Item standard deviations**. The average standard deviation of item scores within a survey is remarkably similar across all developmental levels. Thus, the overall shape of the distribution of the 36 scores within a survey exhibits little change for individuals at different centers of gravity.

3. **Test length.** A Cronbach's-Mesbah analysis of the Cronbachs-alpha score for successively fewer test stems shows that, for a half-item test, the reliability is still excellent (.95); that, even with 10 stems, the SCT has a high degree of reliability (.92); and that, at 5 stems, the reliability is acceptable (.85). These results support 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 (i.e., it appears to measure a single latent variable).

5. **Characteristics of and correlations among each item**. To test the general assumption that the SCT test items, which are designed to triangulate (parallax) the construct from many life contexts, we show that measures of difficulty (mean score) and spread (standard deviation) for each test item, are generally equivalent. However, these measures 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 correlated with the other stems and as having the highest standard deviation. We also show a heat map illustrating the 36-by-36 item correlation magnitudes.

6. **Overall test strength**. Rasch and IRT analyses confirm that the test is robust and sound, as a psychometric measurement, across its entire range. Test reliability and coverage is excellent, and the items have good discrimination. Scoring is consistently of high quality across the items, similar to prior Cronbachs-alpha results.

7. **Construct levels of discrimination**. The test easily differentiates the six-person perspectives and, consistent with the model's hierarchical organization of three parameters, has more difficulty differentiating between levels at the granularity of 12 levels. The 12 stages defined by the model are spaced relatively evenly along the spectrum, supporting 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 were for a new method were that it;

1. be compatible 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; and

3. 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., remain consistent with how, for example, a 3.5/Achiever total score is understood in the field).

The methods we evaluated are described in the paper. Ultimately, we chose the method of taking the mean of the top 6 scores in the survey (the top one sixth of the scores if the survey has more or less than 36 items). Using the formula: (1.2*AveOfTop6)-1, we modified this method slightly to produce a value that is, on average, close to the original ogive method. We call this value the *Focal Score*. Stages International will begin to phase in this method alongside the ogive method, which it will eventually replace over the next year or two.

The Focal Score is a continuous value. Among the benefits of this method are the value's resilience to small perturbations or errors around any cutoff boundaries and its extensibility to different test lengths without having to agonize about revising the set of cutoffs. Another benefit of this method is that it does not confuse psychological regressions, or "shadow crashes," with low complexity or maturity, as is the case with the ogive method. Thus, the focal score can be used for scoring children, which O'Fallon has begun to do.

The revised scoring system will show several measures in addition to the Focal Score: The *Average* across all stems; the *Bottom Score* – the average of the bottom six scores, related to a regression or "shadow" score" for adults; and *Spread* – the spread score is the Focal Score minus the Bottom Score, which indicates the range of values in the survey.

The STAGES framework will gradually shift to the new Focal Score method. We remain agnostic on the question of 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; it is 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. STAGES, on the other hand, allows for the creation of valid surveys with new stems and of any length, which require ongoing adjustments to cutoff values. Such adjustments can only be made arbitrarily. Furthermore, automated computer scoring exists for the STAGES model (see www.stagelens.com), which means the SCT will become more practical to use for large-scale research where the validity of methods is critical.

10. Face Validity: STAGES Applications

The "face validity" of a theory, model, or measurement can be argued for by noting the extent of its use. Given its relative recency, STAGES has seen an impressive degree of use and adoption, including use in a number of doctoral dissertations. To date 12 dissertation (or academic thesis) research projects have made significant use of the STAGES model, and several more are planned for the next few years. This is encouraging for a model as new as STAGES. This journal issue contains papers describing doctoral research projects, including the following.

Dissertations and theses. The "Introduction" article to this journal issue contains a paragraph describing these eight papers appearing in the issue:

- **Abigail Lynam**, *Principles and Practices for Developmentally Aware Teaching and Mentoring in Higher Education.* Lessons learned following dissertation research.

- **Gail Hochachka,** *The scenic route: A developmental approach emphasizes the importance of human interiority in transformative approaches to climate change*. Based on PhD research in progress.

- **Natasha Mantler,** *Women's Authentic Leadership Development.* Based on a completed PhD project.

- **John Churchill** (with Tom Murray) *Integrating Adult Developmental and Metacognitive Theory with Indo-Tibetan Contemplative Essence Psychology.* Based on a completed PhD project.

- **Antoinette Braks,** *Transformational Executive Coaching: The Dimensions, Dynamics, and Dilemmas that Expedite Later Stage Postconventional Leadership Development from Achiever to Strategist.* Base on completed PhD research.

- **Jason Miller,** *Finding Truth Within: Exploring the Importance of Reflective Practice in Deepening Self-Knowledge*. Based on a completed master's thesis.

- **Lisa Buckley,** *Hope Examined Through a Developmental Stage Perspective*. Summary of a PhD research proposal.

The following five dissertation and research projects, which do not have representing articles in this journal issue, also use the STAGES model or assessment. I give longer descriptions of these projects here because they are not described in this issue's "Introduction" article.

- **Eric Reynolds,** *Next-Stage Organizations: A Transdisciplinary Case Study*. The goal of this research was to study the relationship between individual and organizational development and transformation by ascertaining how a Founder/CEO's development informs that of their organization – particularly in leaders or organizations characterized by the post-formal/post-conventional logics that seem necessary to navigate the complexities of contemporary leadership challenges. The study focused on three participants from the sample of CEO's reported on in Laloux's (2014) book *Reinventing Organizations*, which studied post-formal ("teal") organizations. The integral/transdisciplinary qualitative study included data from the STAGES developmental assessment, structured interviews, and a developmental analysis of content from organizational websites and other resources describing the organization. The study included two elements of the STAGES framework: developmental stage, and the epistemic "zone" of the content (from Wilber's 8 zones defined by individual/collective, subjective/objective, first/third person methods). Results of the multi-case study analysis indicate a direct relationship between the Founder/CEO development (and beliefs system) and those of the organization, corroborated Laloux's assertion that his research is a composite representation of a organizations operating at the 4.5 level. Though conclusions from an N=3 case study are provisional, a major contribution of the research was the development of a new transdisciplinary method for assessing organizational development.

- **Jani Attebery**, *Regenerating Soil, Soul, and Society: Garden-Based Sustainability Pedagogy for Incarcerated Adult Learners.* (Completed PhD research). An exploration of the lived experience of prisoners participating in a Sustainable Agriculture Food Production program located within the Arizona State Prison. The mixed-methods study of 10 subjects included the STAGES assessment, a survey assessing past and present experiences of farming and gardening, semi-structured interviews, narrative journals, and observations. The STAGES developmental model functioned as a tool to interpret inmates' experiences. Attebery used the model to describe the merits of, and suggest improvement to, these types of programs by providing flexible educational experiences that were sensitive to the developmental levels of participants.

- **Jimmy Parker,** *STAGES of Organizational Development*. This work reflects on the state of the art in assessing development at the organizational level, and describes some advantages that the STAGES assessment has in that domain. Parker is working on a PhD research project that will apply these ideas to the assessment of development at the group level.

- **Ron Hurst**, *Merging Characteristics of Median Stage Ego Development: When Does Self Initiated Reflection Begin?* (PhD dissertation proposal). A sample of 30 early career professionals in a Southern California distribution company will be given the STAGES assessment. From these, three individuals at each of the Conformist, Expert, and Achiever stages will be selected. Structured narrative interviewing will be used to ask subjects to

reflect upon one or more critical career incidents – potentially life- or career-changing events. Transcriptions of subjects' reflections on how the incident came to be, was or was not resolved, and how it influenced them will be analyzed for developmentally indicative langue and conversational patterns based on O'Fallon's STAGES model and Cook-Greuter's ego development model.

- **Steve Schapiro and Abigail Lynam**, *Transformations in ego development and intercultural sensitivity in graduate students.* This is a soon-to-begin longitudinal study of PhD students in a Human Development and Organizational Development program at a US graduate school. The STAGES assessment, an Intercultural Development Inventory, interview methods, reflective journaling, focus groups, and surveys will be used to study incoming students in pre- and post-assessments as they progress through the graduate program. The goal is to learn what best supports meaning-making development and intercultural sensitivity within the context of graduate study. The study will investigate how the developmental level of the student affects learning and how learning opportunities are utilized. The study will also examine the relationship between intercultural sensitivity and ego development to assess whether intercultural sensitivity is understood and practiced differently at different stages.

Specialty Protocols for understanding life domains. The section above, *Modifications to Test Length and Sentence Starters: Internal Consistency*, describes a number of "specialty protocols" that have been developed based on the STAGES model. We have mentioned six specialty inventories that have had their reliability validated psychometrically on leadership and organizations, love, education, self-understanding, climate change, and a children's SCTs as well as several in-development inventories related to relationships, religious beliefs, spiritual growth, money, hope, dementia, ethics, and parenting.

We have mentioned Hochochka's research using the climate change inventory above. Research is also being planned on how developmental levels differentially create meaning in the areas of money, interpretations of spiritual growth, hope, ethics, relationships, dementia, and love.

Non-research applications. Over 1500 professionals, including many coaches, therapists, organizational consultants, and educators, have taken workshops or training programs through STAGES International on aspects of the STAGES model, and dozens of these professionals are actively using the model in their work. Notable projects include the following:

- **John Kesler** is a theorist and activist in the field of civic and political civility who incorporates perennial spirituality, transpersonal development, and integral theory into his work. He developed and teacher (with colleague Tom McConkie) the "Integral Polarity Practice" (IPP), a transformational contemplative dialogic practice. IPP, which Kesler and McConkie have been teaching to hundreds of individuals for over a decade, makes explicit use of the STAGES model. (See the related article in this journal issue).

- **Marj Britt** is a Senior Minister Emeritus in the Unity Church; and is an author, lecturer, and workshop leader affiliated with the Called by Love Institute. Her work focuses on the

"soul's journey through life" as a "cosmic and human love story." Britt has incorporated the STAGES developmental model into her ministry, and she initiated the specialty inventory on love, which will be used in research projects in the future.

- For 15 years, **Geoff Fitch** and **Abigail Lynam**, along with associates at Pacific Integral, have been running a 9-month cohort program called GTC. GTC makes explicit use of the STAGES developmental model in the design of curriculum and activities, and it teaches the basics of the theory to its participants.

Automated AI-based scoring and large-scale studies. Data from STAGES surveys have been used to create the first automated assessment tool for ego development and meaning-making maturity. As described at www.stagelens.com, the technology is based on a machine learning (artificial intelligence) algorithm that processed approximately 36,000 human-scored sentence completions to build a computer model for predicting developmental level based on sentence completion text. Its accuracy is less than human-scored methods – inventory-level root-mean-square error is estimated at 0.5 developmental levels vs. human checking. It cannot be used to give sufficiently accurate individual scores but is applicable to scaled-up studies that look for statistical patterns aggregated over groups such as pre-post intervention assessments and between-group differences. Several such studies are being planned, and unpublished pilot studies have shown the method to be effective in assessing pre-post gains for educational and transformational interventions.

11. Vocabulary Analysis

[This research has not been reported in any other publications.]

We have also been analyzing the frequencies with which words appear as development progresses through the stages. This produces a rich store of data that we have only begun to take advantage of. Figure 3 below shows a sample for eight different words.

These charts illustrate the relative frequencies of words across the stages. For example, for the word people: the graph title shows that it occurs in 0.457% of all completions in this data sample (this is from a large data sample, but is not our full up-to-date data sample, so the figure is just illustrative). The histogram shows, for each level, the relative percentage of completions at that level that contain "people." Note that it is only the shape of the graphs that can be compared, but not their height. E.G. "space" occurs much less frequently in the data sample (0.05% of completions), so a tall line in that chart cannot be compared to a tall line in the "people" chart. This data is from a sample of 995 surveys with a total of 35,820 completions; containing 1,160,885 total words; and 24491 unique words (ignoring common words like "the," often called "stop words"). The charts does not include stages 1.0 or 6.5, because there are much fewer completions at these levels, and they produce erratic patterns in the graphs.[11] (Ignore the colors in these charts, they are an artifact of the software package used to make the graphs).

[11] E.g. if there are only 10 6.5 completions, and the word "farm" appears in both, it was in a whopping 20% of all completions; but this may be so large that the bar for 6.5 overshadows the rest of the bars, which look so tiny that one can't make out the real pattern.

Figure 3. Sample Word Frequencies.

Love is one of the most common words in the database. You can see its relative frequency gradually increases, with a notable spike at 2.5 and 6.0. Family shows a gradual decrease, with a spike at 2.5/Conventional, which conforms to our understanding of that developmental level (i.e. the importance of conventional in-groups). Note that the STAGES scoring method does not look at specific vocabulary words, but scores on the three parameters: Concrete/Subtle/Metaware, Individual/Collective, and Active/Passive. "Family" is a concrete and collective word, so one might expect it to be prominent at 2nd person perspective (2.0 and 2.5). Note that a completion is scored based on its most developmentally advanced sentence or phrase – the phrase or sentence with "family" in it may be ignored when scoring, if there are parts of the completion that rate higher.

Both "work" and "try" exhibit the pattern that they are more prominent in active (X.5) than in passive (X.0) levels. This also conforms to what we might expect – the active/passive determination is based on verbs and prepositions, and these words would usually appear in an active phrase, especially when they appear as verbs. They spike at the more "conventional" levels of 2.5 and 3.5.

Love has an interesting pattern of spiking at 2.5 and 6.0, suggesting that these are two senses of the word (a conventional sense and a transpersonal sense). Purpose is interesting in that it does not show up at all (or negligibly) until it leaps up at 3.5/Achiever, which conforms to our understanding of that level.

Sense and space were chosen as words that gradually ramp up. One can see that at 5th PP especially, but also at 4th PP, one starts to be aware of subtle "presence," and phrases like "I sense that..." begin to appear. One's sense of time and space are said to alter in the Metaware tier, and we can see an interesting spike in "space" at 6.0.

These words/charts were chosen for illustrative purposes only, to make comments about the shape of the chart – but not to imply anything about the importance of that word related to other words (that requires additional types of statistical analysis, such as the "tf-idf" term-frequency inverse-document-frequency metric). These words were chosen because they coincide with our intuitions about these levels, and seem to confirm the theory. But many words do not have such striking patterns – again our purpose here is to illustrate this type of analysis, not to draw conclusions. Many of the most frequent words had more "boring patterns" that neither agreed nor disagreed with expectations. There were only a very few words that had unexpected patterns that seemed to contradict what one would expect based on how we understand each level. All of this will be reported on in more depth in a future paper.

As mentioned, we have only begun to use this analysis for specific purposes. We can use it in the computer-based automatic scoring project (www.stagelens.com). It could also be used within an exploratory or qualitative analysis of a particular life-domain, such as love; or in characterizing certain levels, such as 5th PP. O'Fallon is also using such analysis to find evidence for different types of state experiences at different levels (see O'Fallon's paper in this issue on States and Stages).

O'Fallon has also been using this type of vocabulary analysis as a cross-checking method during scoring. The main scoring procedure for scoring each completion, as we said, does not mention specific words or themes (this contrasts with the Loevinger scoring method). But once the total protocol score is determined (over the 36 completions), we can use vocabulary as a rubric to double-check the accuracy. Stems are scored independently (locally, not considering the whole inventory). The survey scoring procedure does allow for the scorer to make a judgment call to tweak the total score a bit up or down based on global considerations (especially if it is near a stage border-line). (This final adjustment allowance exists for all scoring methods in the Loevinger tradition. The adjusted score is what is given to clients, but note that it is only the pre-adjusted score that is used for all quantitative research studies). O'Fallon has developed a "rubric" method that shows common words from all four "quadrants" that appear at each level (and don't tend to appear before then). If an inventory does not have a sufficient representation of such words, the analyst is encouraged to double-check the accuracy of the final (adjusted) score.

12. Studies of Children's Inventories

[This section summarizes research reported in "Children's Inventories" by O'Fallon, in this issue]

Adjusting stage calculation to work for children: the Core Score method. As described in the paper "Investigating the Validity of the Ogive Aggregation Method..." (in this issue and summarized in Section 9 above) the ogive method currently used to aggregate the individual item scores confuses psychological regressions ("shadow crashes") with low complexity/maturity – i.e. momentary regressions to lower ego levels that may be triggered based on the subject of the stem. In Section 9 we described a new aggregation method called the "focal score" that does not have this issue. This allows the sentence completion instrument to be used for assessing children, and O'Fallon has begun a research study on just that.

More specifically, the body of research based on Loevinger's SCT was largely limited to adults, but included some teenagers. It included some studies of prison populations and people of lower socioeconomic and/or educational levels. Those exhibiting very low developmental levels, i.e. 1.0, 1.5, and 2.0, are demonstrating levels of meaning making complexity or self-understanding normally attributed to young children. (We assume, as prior researchers did, that test subjects exhibiting lower developmental levels did not have genetic or neurological conditions leading to intellectual disabilities, but that their lower-than-the-norm scores have developmental or psychological origins). The scoring systems used by WUSCT and MAP are geared to adult (and young adult or teen) populations, and one will get a 1.0/Impulsive or 1.5/Opportunist (i.e. 1st PP) completion score by showing regressive behavior, including impulsive outbursts with vulgarity or violence (e.g. "sucks," "I hate you"), or responding with single concrete words or phrases (e.g. Raising a family…"family"). (Note that these categories do not exhaust the possible 1st PP territory).

The ogive aggregation rules, inspired by Bayesian statistics, prioritize rare behaviors and require only 7 of the 36 completions at 1.0, 1.5, or 2.0 to score an entire protocol ("center of gravity") at that level (see "Investigating the Validity of the Ogive Aggregation Method..."). Getting such a low rating can come from a very simplistic answer, but it can also come from what seems a narcissistic or impulsive answer. Thus we argue that, not only the item-scoring rules, but more importantly the ogive aggregation 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 get an ogive score of 2.0/Rule-based if they have only *7 stems* at 2.0 or lower. Clearly such a person's "center of gravity" should be more like 3.5 than 2.0. Because there are very few people with a center of gravity below 2.5 in the populations that currently use the SCT, this problem has not been particularly salient. But one area where it is particularly limiting is in studying children.

Other developmental assessments, such as those based on hierarchical complexity theory, do not have this limitation. Because the STAGES scoring method uses language properties rather than exemplars, it is not susceptible to this problem in scoring items; but, as long as it uses the ogive cutoff method, it is susceptible to this distortion at the whole survey level. Thus, as mentioned, we are moving to an alternate aggregation method called "focal score" that is more like an average.

First assessment of children using the SCT. With these biases in the scoring system removed, the STAGES model can be used to assess developmental levels and developmental phenomena in younger children. As mentioned above in Section 5, and detailed in the paper (in this issue) "The STAGES Specialty Inventories..." O'Fallon and colleagues have developed and validated a "specialty inventory" with sentence stems modified to work for children (the list of stems is given in that paper). The children's inventory, for pre-school through junior high school, was co-created by Jennifer Haynes, Kim Barta and Terri O'Fallon. It is meant to be administered orally and transcribed, especially to children in grade school.

In a recent study 53 children, age range 4-13 enrolled in a progressive elementary school in Brisbane Australia, gave verbal answers to in-person prompts; recordings of their answers were

transcribed. The paper "STAGES Child Research: Preliminary Report" by O'Fallon, in this issue, summarizes some preliminary results. It is the first study applying the sentence completion method to young children (performed using transcriptions from face-to-face verbal interviews).

11 new sentence starters were created to make a 36-item SCT suitable for children. Cronbach's Alpha test for internal consistency indicated that the test overall was valid (alpha=0.88, "good," almost "excellent"), and that the new items by themselves had only a "questionable" alpha (0.63). We will look into what this tells us about improving the new stems, but overall the test is valid enough to measure the development of the target audience.

The test was administered verbally to 53 children at their school in Brisbane, and the audio was transcribed. An expert scorer scored the results, which are shown in Figure 4 as frequencies in histograms at both the item level and the survey level.

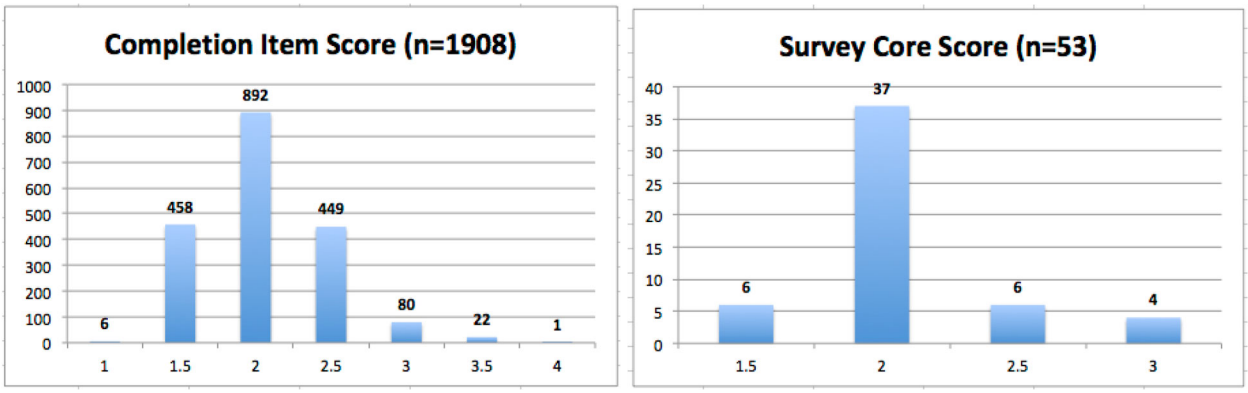

Figure 4. Child Study results: frequency histograms.

From the Survey graph we can note that the vast majority of children scored at 2.0, with a few at 1.5, 2.5, and 3.0. From the completions graph we can see that there are also a few item scores as low as 1.0, and as high as 4.0

O'Fallon continues to analyze this data to derive descriptive accounts of how the children make meaning of various themes (to be published later).

13. Conclusions

We have summarized research projects that have shown substantial evidence of the validity and usefulness of the STAGES model. This research includes;

- a summary of **prior research** on the SCT method, which establishes a baseline for STAGES validity;

- a **replication study** showing that STAGES tracks well with the MAP method, from which it was derived, up to 4.5/Strategist, which establishes the STAGES scoring method solidly within the Loevinger SCT tradition;

- studies showing that the STAGES assessment remains robust for **variations in sentence stems and test lengths** (including the design of specialty inventories), which will allow for wider application of the ego development construct;

- **longitudinal** data analysis that shows, among other things, that the MetAware (Tier 3) stages show the appropriate developmental progression, which is important because the studies mentioned in #2 did not thoroughly describe Tier 3 validity;

- additional analysis of **late stage** sentence protocols, checking that metaware protocols contain a reasonable number of complex 4.5/Strategist item scores;

- additional **inter-rater** analysis, including recently trained scorers and per-item IRR statistics, has shown exceptionally high inter-rater reliability, even at the item level;

- a research study that used **item response theory and Rasch analysis** to investigate many psychometric properties of the STAGES assessment and that also elaborated on a number of problems with the ogive item-aggregation method, suggesting an alternative. This modification allows us to better use the assessment for (1) children and (2) shadow evidence (vs. prior SCTs);

- a "Face Validity" section describing various uses of the STAGES model, including its use in 10 dissertation or thesis projects, and an indication of many projects to come;

- **vocabulary analysis** showing promising analytical methods for (1) more detailed characterization of each level, and (2) refining the scoring system; and

- studies of **children's item responses** that illustrate the first application of the SCT (that we are aware of) to populations in the grade-school level.

STAGES is a new developmental model and is sure to evolve as new empirical research and theoretical critiques accumulate in the future. It has established a strong foundation upon the Loevinger lineage of SCT assessments, which is quite robust, and our STAGES research is already producing results that inform all variations of the SCT.

Some of the more controversial issues surrounding the STAGES model include:

- characterizing development in the highest levels, 5.0/Construct Aware and above (Alternate theories exist in this territory, and little to no empirical work has been done to compare them).;

- the comparative validity and usefulness of "wide" or "holistic" developmental constructs, such as ego development or Kegan's levels of consciousness, vs. narrower constructs, such as reflective abstraction, self-understanding, relationship sophistication, and leadership maturity (See Jordan and Murray's article in this issue for more on this theme).;

- the definition of states of consciousness and the relationship between states and stages – O'Fallon has a particular theory of this worked into the STAGES framework, see her article in this issue, but it remains an active area of debate in the field; and

- over-fitting of developmental models – developmental models afford the categorization of people into developmental levels, which, while especially useful, also risks the reductionism of using a single term to describe the complexity of human cognition and personality. (Though all the major developmental theorists warn against this, there are insufficient research and methodology best practices to support a rich articulation of individual differences that can enrich the description of a person's developmental profile).

In addition, an important open question for all developmental theories is how developmental phenomena apply to collectives (i.e., relationships, groups, organizations, and populations). Many non-validated theories and several small empirical studies exist, but this is a key emerging in need of additional work.

Future research comparing developmental models. Taking a larger perspective, we can compare the Loevinger SCT lineage to other developmental assessments to suggest important work yet to be done. Though many developmental models and psychometrically validated assessments occupy the long scholarly history since Piaget's early work, only a small set of these seem relevant for those attempting to put this research into practice. The first is Robert Kegan's construct developmental model, with its "Subject-Object Interview" assessment method. This process is judged by many, including Kegan, to be measuring a construct (e.g., meaning-making maturity or order of consciousness) that is closely related to the construct of "ego development" measured by the SCT. However, we know of no empirical work that investigates their correlation. Second are the two similar neo-Piagetian developmental theories by Michael Commons and colleagues (MHC, Commons et al., 1988), and Kurt Fischer and colleagues (Skill Theory, Fischer, 1980). There appears to be no published work that thoroughly compares and contrasts these frameworks from either an empirical or theoretical perspective, though some preliminary unpublished studies do exist. See Murray (2009) for preliminary musings. The theoretical foundations of these two lineages (Loevinger/Kegan vs. the neo-Piagetian) suggest that they should have important differences and useful similarities – all of which will hopefully be explored in future research.

References

Aurobindo, S. (1992). *The synthesis of yoga*. Pondicherry, India: Sri Aurobindo Publication Department.

Barta, K. (2020). Psychological Application of the STAGES Model. *Integral Review*. This issue.

Blumentritt, T. (2011). Is higher better? a review and analysis of the correlates of post-conventional ego development. In *Pfaffenberger, A. Marko, P., & Combs, A. (Eds)., The Post-conventional Personality: Assessing, Researching, and Theorizing Higher Development*, 153-162.

Browning, D. L. (1987). Ego development, authoritarianism, and social status: An investigation of the incremental validity of Loevinger's sentence completion test (short form). *Journal of Personality and Social Psychology*, *53*(1), 113–118.

Carlson, V.K., & Westenberg, P.M. (1998). Cross-cultural research with the WUSCT. In J. Loevinger (Ed)., *Technical foundations for measuring ego development* (pp. 57-75). Mahwah, NJ: Erlbaum.

Cohen, J. (1968). Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin, 70*, 213–220.

Cohn, L. D., & Westenberg, P. M. (2004). Intelligence and Maturity: Meta-Analytic Evidence for the Incremental and Discriminant Validity of Loevinger's Measure of Ego Development. *Journal of Personality and Social Psychology*, *86*(5), 760-772.

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. http://www.cook-greuter.com/Cook-Greuter%209%20levels%20paper%20new%201.1'14%2097p%5B1%5D.pdf

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

Feinstein, A. R., & Cicchetti, D. V. (1990). High agreement but low kappa: I. The problems of two paradoxes. *Journal of clinical epidemiology*, *43*(6), 543-549.

Fischer, K. (1980). A theory of cognitive development: The control and construction of hierarchies of skills. *Psychological Review, 87*(6), 477-531.

Forman, M. (2010). *A guide to integral psychotherapy: Complexity integration and spirituality in practice*. Albany New York: Suny.

Harris, S. (2005). *The end of faith: Religion, terror, and the future of reason*. New York: WW Norton & Company.

Holt, 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.

Hy, L. X., & Loevinger, J. (1989). *Measuring ego development*. London: Psychology Press.

Jespersen, K., Kroger, J., & Martinussen, M. (2013). Identity status and ego development: A meta-analysis. *Identity*, *13*(3), 228–241.

Kegan, R. (1994). *In over our heads: The mental demands of modern life*. Cambridge, MA: Harvard University Press.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33* (1), 159–174.

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

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

Manners, J., & Durkin, K. (2001). A critical review of the validity of ego development theory and its measurement. *Journal of personality assessment*, *77*(3), 541-567.

McCauley, C. D., Drath, W. H., Palus, C. J., O'Connor, P. M., & Baker, B. A. (2006). The use of constructive-developmental theory to advance the understanding of leadership. *The Leadership Quarterly*, *17*(6), 634-653.

Murray, T. (2020) Investigating the Validity of the Ogive Aggregation Method, Including the use of Rasch Analysis, for the Sentence Completion Test and the STAGES Model. *Integral Review*. This issue.

Murray, T. (2009). Intuiting the Cognitive Line in Developmental Assessment: Do Heart and Ego Develop Through Hierarchical Integration? *Integral Review*, Vol. 5 No. 2, December 2009, p. 343-354.

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

Novy, D. M., & Francis, D. J. (1992). Psychometric properties of the Washington University Sentence Completion Test. *Educational and psychological measurement*, *52*(4), 1029-1039.

O'Fallon, T. (2010). *The collapse of the Wilber-Combs Matrix: The interpenetration of the state and structure stages.* Presented at the July 2010 Integral Theory Conference.

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

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

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

O'Fallon, T., Polister, N., Blazej Neradiek, M., Murray, T. (2020). The Validation of a New Scoring Method for Assessing Ego Development Based on Four Dimensions of Language. *Heliyon* 6(3), E03472, March 01, 2020; https://doi.org/10.1016/j.heliyon.2020.e03472.

O'Fallon, T. & Murray, T. (2020). The STAGES Specialty Inventories: Robustness to Variations in Sentence Stems *Integral Review*. This issue.

Pfaffenberger, A. H., Marko, P. W., & Combs, A. (Eds). (2011). *The postconventional personality: Assessing, researching, and theorizing higher development*. SUNY Press.

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

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

Westenberg, M. P., Drewes, M. J., Goedhart, A. W., Siebelink, B. M., & Treffers, P. D. (2004a). A developmental analysis of self-reported fears in late childhood through mid-adolescence: social-evaluative fears on the rise? *Journal of Child Psychology and Psychiatry*, *45*(3), 481-495.

Westenberg, P. M., Hauser, S. T., & Cohn, L. D. (2004b). Sentence completion measurement of psychosocial maturity. *Comprehensive handbook of psychological assessment*, *2*, 595-616.

Wilber, K. (1995), *Sex, Ecology, Spirituality: The Spirit of Evolution*. Boston: Shambhala.