The STAGES Specialty Inventories: Robustness to Variations in Sentence Stems

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

Abstract: The STAGES developmental model is a variation of prior ego development frameworks that defines developmental levels in terms of three parameters: object of awareness (concrete, subtle, or metaware), individual vs. collective focus, and active/passive orientation. STAGES, like prior frameworks, uses a sentence completion test (SCT) assessment. Prior frameworks rely on exemplar-based scoring that is closely tied into the specific sentence stems. In contrast, STAGES scoring system is based on language properties that do not depend on the sentence stem. Thus STAGES is the first such assessment to be able to freely incorporate alternative sentence stems without the labor intensive process of discovering the full range of specific responses. Though the STAGES theory and assessment methodology easily allow for using alternative sentence stems, the validity of using alternative stems needs to be shown. In this paper we report on internal consistency studies of several "specialty protocols" which are SCT surveys with 6 to 10 of the original 36 stems replaced by stems focused on a particular specially area. Results show strong reliability scores, via the Cronbach's alpha statistic, for six specialty inventories, on: leadership and organizations, love, education, psychological reflection, climate change, and a children's SCT.

Keywords: Cronbach's alpha validity, children's and adult assessments, ego development, protocol variations, STAGES.

Introduction and Background

Motivations. The STAGES developmental model is a variation of prior ego development frameworks that defines developmental levels in terms of three "parameters" (or "drivers"): object of awareness (concrete, subtle, or metaware), individual vs. collective focus, and active/passive orientation. It uses a variation of the sentence completion test (SCT) that began

[1] **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 Masters degrees in Special Education and Spiritual direction, and has an Integral PhD in Transformative Learning and Change.
terri@stagesinternational.com

[2] **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

with Loevinger's Ego Development research (1998) and was updated by Cook-Greuter (1999). The prior frameworks rely on exemplar-based scoring that is closely tied into the specific sentence stems. In contrast, STAGES scoring system is based on language properties that do not depend on the sentence stem. Thus STAGES is the first such assessment to be able to freely incorporate alternative sentence stems without the labor-intensive process of discovering the full range of specific responses. Though the STAGES theory and assessment methodology easily allow for using alternative sentence stems, the psychometric validity of each use of alternative stems still needs to be shown. In this paper we report on internal consistency studies of several "specialty protocols" which are SCT surveys with 6 to 10 of the original 36 stems that are replaced by stems focused on a particular specially area. We will also summarize research on the validity of SCTs of varying length, another dimension of survey variability.

Prior Research

Validity of the SCT and the STAGES scoring system. The SCT for ego development is a robust assessment instrument. Westenberg et al. (2004) conclude that "findings of over 350 empirical studies generally support critical assumptions underlying the ego development construct" (p. 485). In Murray (2017) we give an overview of the 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. A summary of those findings is also included in *A Summary of Research on and with the STAGES Developmental Model* in this issue. In that article, we summarize validity studies particular to the STAGES model. In this paper we are interested in two issues: the primary one being the robustness of the SCT model to variations in sentence stems, and secondarily, its robustness to reducing the number of sentence stems. Next we summarize prior research on these two questions.

Variations in SCT stems. Historically, within the lineage of ego development research (and several other developmental models), stages of human development have been identified by interviewing or observing samples of people, recording their responses and organizing these responses into categories. Researchers then sequenced these categories of responses from the earliest to the latest level of maturity. These categories could then be used to identify where a person might fall in their developmental journey through life. This is the method used by other SCT projects and is outlined in a dissertation on the construction of and validity measures of one new sentence stem "A good Boss" (Minard, 2009).

As to variations in the choice of sentence stems ("test items"), we can make several observations. Loevinger adapted the sentence items numerous times before settling on a final version. There is nothing particularly special about the stems Loevinger used. Though they were carefully chosen and vetted; many were drawn from the experience of prior researchers, and the entire set evolved over the years before settling into the current standard form of the WUSCT. Others have used alternate forms of the WUSCT tailored to men, women, or youth.

Sentence starters can be inadequate for a variety of reasons, e.g. they may be vague and thus understood in very different ways; they may coerce an overly limited developmental range of responses; or they may introduce biases of various types. Thus new stems must be pilot tested for

clarity and psychometric validity. But assuming that this due-diligence work is done to ensure that questions are adequate, overall prior research *suggests* that the strong psychometric properties of the SCT are robust to changes in the choice of sentence stems. However, the general question of stem flexibility has not been addressed until this study. Rather, specific sets of stems have been tested in the past.

Ego development (meaning-making complexity) is a holistic capacity spanning all life-contexts (though we may exhibit maturity different than our "center of gravity" in any given context). The stems are meant to probe across a variety of contexts, and triangulate toward an overall measurement. Cook-Greuter (1999) notes that:

> As Fischer, Hand, & Russell (1984) pointed out people tend to respond optimally to a task in 'domains in which they are highly motivated.' To tap this motivation, the SCT stems were devised to address ordinary everyday experiences shared across a wide spectrum of people. (p. 52)

Kegan's and Wilber's framing of a holistic span of life contexts as including subjective, intersubjective, and objective (I/we/it) contexts. Loevinger (1985) describes the span of stems in a related way:

> Looking at item content, the stems can be classed as *first* person (My father –, When they talked about sex, I –), *third* person (Sometimes she wished that –, Usually he felt that sex –), and common noun or *impersonal* (A good father –, Being with other people –). (p. 424)

Some stems seem to be sensitive to particular transitions along the developmental spectrum, and this is another reason for having an adequate diversity of stems. For example, some stems are related to impulse control which comes on line at second person perspective, and increases gradually or levels off for later levels. Therefore, sentences sensitive to impulse control are also sensitive to the transition from pre-conventional to conventional levels. Abstract and formal thinking begins at third person perspective, and increases gradually or levels off after that, so we would expect that certain stems are more sensitive to transitions in this part of the spectrum. In general we would expect that, to a weak but statistically significant degree, certain sentence stems are better at signaling changes in specific levels. Murray (2020) includes item response theory analysis that shows which stems do in fact load for higher vs. lower levels (i.e. "item difficulty"). However, in general, the data indicate that all of the stems are of comparable difficulty, and that every stem elicits responses from across the developmental spectrum, as is intended by the SCT design.

Torbert & Livne-Tarandach (2009) report on variations of the WUSCT that have evolved into Cook-Greuter's MAP and Torbert's LDP and GLP instruments. Cook-Greuter and Torbert modified Loevinger's protocol to "omit a number of gender-based items and, include work and leadership-related stems" (IBID, p. 132). Torbert's LDP is also shorter (24 items) and includes six new stems not in the WSUTC. Torbert & Livne-Tarandach (IBID, p. 134) report that "the responses to the new stems correlate better with an individual's overall profile rating than responses to the former stems did, thus improving the overall reliability of the measure."

O'Fallon made additional adjustments to stems to derive her General protocol (or "inventory"), the standard set used in her research and consulting before she began branching out into "specially protocols." As described below, modifications were to insure that that stem collection had an equal representation from the "four quadrants" of I/we/it/its as explained later.

Variations in SCT Length. As to the *length* of the SCT: Novy & Francis (1992) compared the split-half versions of the WUSCT (18 items each). They conclude "these results provide empirical justification for those users of the SCT who have the need for shorter, interchangeable, and reliable forms of the test."[3] Holt (1980, p. 909), experimenting with a 12-item short form of the WUSCT found that inter-rate reliability was "at least as good as…reported by Loevinger; and the internal consistency…was quite adequate....Analyses of other data indicate that the short forms are representative samples of the full [WUSCT]".[4] He goes on to say the data show that "an abbreviated form of Loevinger's WUSCT is a reasonably reliable, feasible, and useful instrument for large-scale research...[and is]… a representative sample of the larger instrument, which probably gives substantially the same results" (p. 916). Basic psychometric theory predicts that more evidence will result in better accuracy, so for individually-based assessments, done for coaching or consulting purposes, the full set of items is still recommended, but for research or group-statistical assessments, it would appear that shorter forms are quite valid.[5] New research related to inventory length, analyzed for STAGES data, is summarized later.

The STAGES Matrix

The STAGES Matrix is a new model that describes the stages of human perspectives from birth to the latest levels of human development investigated by research. One of the innovations of STAGES is the use of parameters for describing developmental stages. These are underlying attributes that lead to development. Below is the STAGES MATRIX which shows the sequence of patterning of the parameters of human consciousness across 12 levels (or 6 person perspectives).[6] These parameters are arrived at by asking 3 fundamental questions, as shown in Figure 1. We refer the reader to other papers, including Barta's "Psychological Application of the STAGES Model" in this issue, for a description of The STAGES model.

The STAGES General Protocol: Stems and Quadrants

Above we mentioned the notion, articulated by Kegan and Loevinger, that the holistic triangulation of contexts needed to measure one's general "center of gravity" should include subjective, objective, and intersubjective (I/we/it) contexts. O'Fallon has extended this notion to

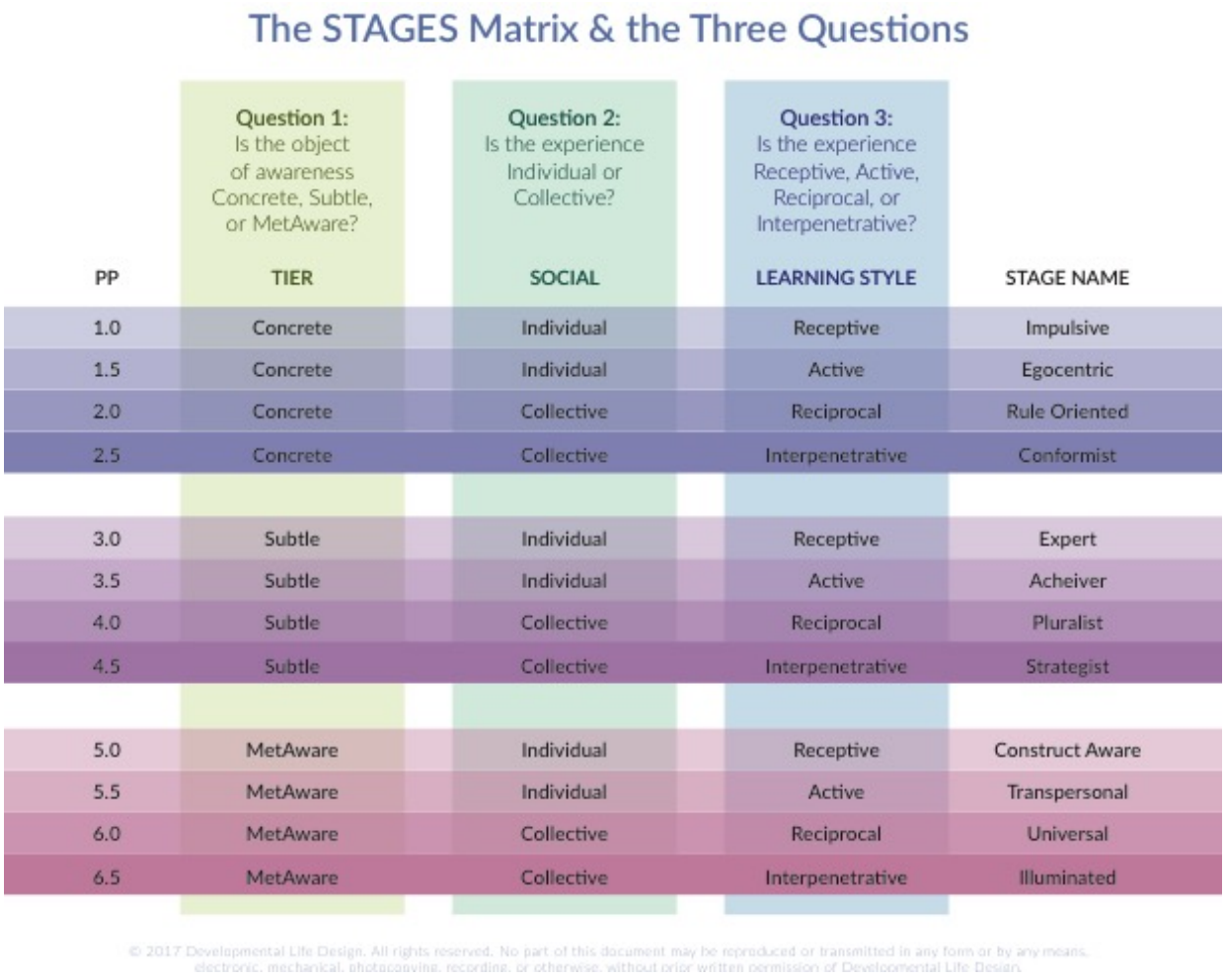
[3] Westenberg (2004, p. 603) claims that "Several studies suggest that the split-half reliability of the WUSCT…is about .80, and, if disattenuated for the greater unreliability of the two test halves, the correlation between the two halves approached unity."

[4] Browning (1987) used a short, 12-item, form in her study, and found that the short form has inter-rater agreement and internal consistency (alpha) "comparable to those reported by Loevinger" (p. 114).

[5] Holt (1980) also reports that the Cronbach's alpha for the 12-item test was .77, vs. Loevinger's finding for .91 for all 36 items, which is exactly what psychometric theory would predict for a randomly selected subset of 12 items from 36 (p. 915).

[6] The full model includes a third tier for a total of 16 levels, but the third tier is hypothetical and not supported by data.

use Wilber's four-quadrant model: I/we/it/its, where the objective domain has been split into individual and collective objects (matching the subjective domain's split between individual "I" and collective "we").[7] Thus O'Fallon's protocols are designed to have a relative balance of stems representing each of the four quadrants. The quadrant categorization of stems is approximate, and is not expected to have a strong influence on the nature of completions – i.e. and subjects can respond to any of the stems in ways that emphasize any of the quadrants.[8] (Preliminary research is inconclusive RE whether stems categorized in a given quadrant tend to elicit responses from that quadrant – more research is needed.)



The STAGES Matrix & the Three Questions

PP	Question 1: Is the object of awareness Concrete, Subtle, or MetAware? TIER	Question 2: Is the experience Individual or Collective? SOCIAL	Question 3: Is the experience Receptive, Active, Reciprocal, or Interpenetrative? LEARNING STYLE	STAGE NAME
1.0	Concrete	Individual	Receptive	Impulsive
1.5	Concrete	Individual	Active	Egocentric
2.0	Concrete	Collective	Reciprocal	Rule Oriented
2.5	Concrete	Collective	Interpenetrative	Conformist
3.0	Subtle	Individual	Receptive	Expert
3.5	Subtle	Individual	Active	Acheiver
4.0	Subtle	Collective	Reciprocal	Pluralist
4.5	Subtle	Collective	Interpenetrative	Strategist
5.0	MetAware	Individual	Receptive	Construct Aware
5.5	MetAware	Individual	Active	Transpersonal
6.0	MetAware	Collective	Reciprocal	Universal
6.5	MetAware	Collective	Interpenetrative	Illuminated

Figure 1. The STAGES Matrix.

In 2016 O'Fallon revised Cook-Greuter's sentence stems by changing 9 of the original stems to new ones to achieve a relative balance among the quadrants. Adjustments were also made because some of the original Loevinger stems seemed outdated vs. 21st century western cultural norms. Table 1 shows the results. Of the 36 items, there are 8 stems associated with each of the four quadrants, plus 4 stems that are mixed-quadrant. Changed stems are indicated by bold text.

[7] And see a similar 4-domain model by Søren Brier (2013).

[8] The determination of stem quadrant is somewhat subjective, as judged by O'Fallon. We do not describe a method defining each quadrant here.

Table 1. General Inventory: Showing New vs. Prior Sentence Stems (changes in bold).

Order	Old Protocol	New General Protocol	Quadrant
1	Raising a family	Raising a family	LL
2	When I am criticized	When I am criticized	UL
3	Change is	Change is	UR
4	**A man's job**	**These days, work**	**UR**
5	Being with other people	Being with other people	LL
6	The thing I like about myself is	The thing I like about myself is	UL
7	**My mother and I**	**My co-workers and I**	**LL**
8	What gets me into trouble is	What gets me into trouble is	UR/UL
9	Education	Education	LR
10	When people are helpless	When people are helpless	LL
11	**Women are lucky because**	**What I like to do best is**	**UR**
12	A good boss	A good boss	UR
13	**A girl has a right to**	**We could make the world a better place if**	**LR**
14	The past	The past	LR
15	**When they talked about sex, I**	**Privacy**	**UL**
16	I feel sorry	I feel sorry	UL
17	When they avoided me	When they avoided me	LL
18	Rules are	Rules	LR
19	Crime and delinquency could be halted if	Crime and delinquency could be halted if	LR
20	**Men are lucky because**	**Business and society**	**LR**
21	I just can't stand people who	I just can't stand people who	UL
22	At times s/he worried about	At times I worry about	UR
23	I am	I am	UL
24	If I had more money	If I had more money	UR
25	My main problem is	My main problem is	UL
26	When I get mad	When I get mad	UR/UL
27	**People who step out of line at work**	**People who step out of line**	**LL, LR**
28	**A husband has a right to**	**A partner has the right to**	**LL**
29	If my mother	If my mother	LL
30	If I were in charge	If I were in charge	LR
31	My father	My father	LL
32	If I can't get what I want	If I can't get what I want	UR
33	When I am nervous	When I am nervous	UL
34	**For a woman career is**	**Technology**	**LR**
35	My conscience bothers me if	My conscience bothers me if	UR
36	Sometimes s/he wished that	Sometimes I wished that	UL, UR, LL, LR

Specialty Inventories: Description

As mentioned, the STAGES model allows for efficient introduction and experimentation with new sentence starters. O'Fallon and colleagues have developed a number of "specialty protocols." The first six of these, which replace 6 of the General protocol stems with theme-specific stems, have been validated for internal consistency: leadership, love, education, and climate change, psychological reflection, and a children's version of the SCT. Reliability analysis results are in the Results section later.

These spaciality inventories were developed separately and at different times between 2016 and 2020. With each specialty inventory, we worked with a consultant who specialized in that domain. We worked with that consultant to design 6 stems targeted at that domain, and in deciding which of the 6 General inventory stems should be replaced. Through this process we maintained the goal of having adequate stem representation among the four quadrants. Sometimes stems were spot-tested on a small number of subjects to check for potential problems (such as common misunderstandings), before the final set was decided upon.

Why create specialty protocols? We are testing the person perspectives that people take on these different areas. For example, what is a first, second, third, fourth, fifth, sixth person perspective on leadership; on love; on education; on climate change? For example on the Love inventory Britt and O'Fallon were able to categorize the specific ways that people at each person perspective understood the topics of love. In the climate change area this might be helpful in planning, policy decisions, or communications that are meaningful to various audiences. Specialty inventories are created for several reasons:

- Quantitatively, they allow a researcher or leader to gauge meaning-making (developmental) levels for a specific theme, for specific populations or sub-populations;

- Qualitatively, they allow for research into the concepts and themes by which individuals understand some domain, e.g. climate change;

- Finally, they allow for the wording of an assessment to appear more familiar or comfortable for given test-taking audiences. For example, if one is assessing employees in an organizational transformation research project, one may want to remove questions about parents or personal relationships which may strike the subjects as too intrusive.

Table 2 shows the new stems designed for each of the six specialty inventories we validate below: leadership, love, education, psychological reflection, climate change and a child inventory which is given orally. 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.

Table 2. New stems for the specialty inventories.

Inventory	New stem
Leadership	A good organization
Leadership	These days progress means
Leadership	Working with other people
Leadership	We could make organizations better if
Leadership	When developing strategies...
Leadership	These days, a career
Leadership	A good leader
Love	Love changes when
Love	Loving other people
Love	Love
Love	Business, society, and love
Love	When I am in love
Love	My main problem with love is
Education	These days, teaching
Education	My students and I
Education	A good educator
Education	Education and society
Education	Students who step out of line
Education	A teacher has the right to
Climate change	The environment
Climate change	Climate change is
Climate change	Regarding climate change, I
Climate change	My biggest concern about climate change
Climate change	People who deny climate change
Climate change	Actually, climate change...
Psychological	My friends and I
Psychological	A healthy person
Psychological	Sex
Psychological	Psychology and society
Psychological	An intimate partner has the right to
Children's	My family
Children's	Grandparents
Children's	These days, school
Children's	My parents and I
Children's	A good child
Children's	When they didn't let me join in
Children's	Bullying could be stopped if
Children's	Children and parents are lucky when
Children's	Children who step out of line
Children's	A parent has the right to
Children's	A child has the right to

The six validated specialty inventories came about as follows:

- The Leadership/organizational/business inventory was initiated by John Wood and Tamara Androsoff, PhDs, an OD consultant and coach who works with the STAGES model – with the goal of including sufficient stems addressing leadership and organizational issues.

- The Love inventory was pioneered by Marj Britt PhD, who is a Unity minister who became interested in the STAGES developmental model. Marg's specialty is the area of personal and spiritual manifestations of Love, and she wanted to create a Love inventory.

- The Education inventory was initiated by Abigail Lynam, a college professor who teaches courses on educational theory, sustainability, and developmental theory.

- The Climate Change inventory was initiated by Gail Hochachka, a researcher in the area of Climate Change, who used the new Climate Change inventory in her dissertation study (see her report in this journal issue).

- The psychological reflection inventory was initiated by Mark Forman PhD and Kim Barta, MA. Both psychotherapists. It removes stems added to the General protocol about work and society, to maintain a focus on psychological/therapeutic themes.

- The children's inventory was co-created by Jennifer Haynes, Kim Barta and Terri O'Fallon. Jennifer is the principle of the integrally-informed Brisbane Independent School working with children from preschool age through Junior High, approximately. It is mean to be administered orally to children in that age range.

Method: Reliability assessment

As mentioned above, unlike prior models, the STAGES theory and scoring method allows for easy substitution of new stems. However, for any new protocol, one must establish the new stems, each and as a whole protocol, maintain psychometric validity. We assess the reliability of the SCT using the standard Cronbach's Alpha measure of internal consistency (George & Mallery, 2003). A commonly accepted rule of thumb for describing internal consistency is as follows (IBID):

Table 3. Cronbach's Alpha Interpretation.

Cronbach's alpha	Internal consistency
$\alpha \geq 0.9$	Excellent
$0.9 > \alpha \geq 0.8$	Good
$0.8 > \alpha \geq 0.7$	Acceptable
$0.7 > \alpha \geq 0.6$	Questionable
$0.6 > \alpha \geq 0.5$	Poor
$0.5 > \alpha$	Unacceptable

Our statisticians determined that a sample of 20 surveys was sufficient to validate the reliability (internal consistency) of any new protocol.[9] Later we will give the reliability metrics for each of six new inventories (protocols) that have been designed and tested. But first we report on the reliability of the General Protocol, upon which the specialty protocols are based.

Following inventory design, we worked with the specialist to recruit 20 or more volunteers to take the assessment. Diversity of backgrounds and demographic characteristics was intended, but not strictly monitored. Surveys were scored by a certified STAGES scorer using the standard scoring manual/method. Then our statisticians ran the internal consistency statistics on the data set. In all cases, as reported below, the first version of the new inventory was satisfactory – i.e. there were no outlier stems that needed to be re-designed.

In the results for each inventory, we show the Cronbach's alpha for the inventory as a whole, for the inventory without the new stems, and for only the new stems. Note that Cronbach's alpha values are naturally higher when there are more stems, and values for different numbers of stems are not directly comparable. That is, the internal consistency for 6 stems is always lower, but this does not necessarily mean that the items are worse then the other items. One would have to compare Cronbach's alpha values between the same number of items for a direct comparison.

There are two additional types of psychometric analysis possible with each new inventory. One removes each item in turn, and checks the Cronbach's alpha of the remaining set of 35 stems – repeated for each of 9 new stems. We decided to only use this analysis when the overall reliability score is low, to check on outlier stems, but we have not needed to do this thus far. In addition, our statistician reports include a Spearman correlation dendrogram (Sokal & Rohlf,1962) that can be used to reveal hierarchical clustering patterns, i.e. similarities or correlations, between stems. These reveal in general that there are no outlier stems, and we have not yet used these analyses for deeper interpretations, so we do not include them in this paper.

Results

The General Inventory

Table 4 shows results of assessing the internal consistency of the new (now standard in the STAGES assessment) General protocol. It was validated vs. the prior "old" protocol. We compared 940 surveys based on the old protocol with 351 based on the new protocol.

Table 4. Cronbach's Alpha for New (General) vs Old protocol.

	Old *(N=940)*	**New** *(N=351)*
All 36 stems	0.98	0.97
Common stems	0.97	0.96
Changed stems	0.91	0.89

[9] Our statistical consultant for this project is Moni B. Neradilek from The Mountain-Whisper-Light Statistics in Seattle, WA.

In a test of removing each item and checking the Cronbach's alpha of the remaining set of 35 stems, repeated for each of 9 new stems, the resulting value varied from 0.9688 to 0.9692.

In related research reported on in this journal issue ("Investigating the Validity of the Ogive Aggregation Method, Including the use of Rasch Analysis, for the Sentence Completion Test and the STAGES Model") we use item response theory to assess overall reliability of the test, and look for item outliers (see subsections "Test length" "Item difficulties and item standard deviations", "Overall test strength," "Item discrimination," and "Summary/fit table; and discrimination strata"). These results reiterate the high reliability scores for the full inventory, and the acceptability of each individual item, as indicated above. (However, the analysis does show that the "I am" stem stands out as being the least correlated with other items.)

Conclusion. The overall internal consistency of both the old and new general protocols as measured by the Cronbach's alpha statistic was excellent (0.98 and 0.97, respectively). The internal consistency of the changed stems in the new protocol was good and only slightly smaller than the internal consistency of the changed stems in the old protocol (α = 0.89 vs. 0.91). Through the method of deleting each new stem one at a time, we confirm that each new stem does not effect the overall consistency of the protocol.

Specialty Inventories: Reliability

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 New England, gave verbal answers to in-person prompts; recordings of their answers were transcribed. The data is shown in Table 5

Table 5. 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 adults' (see O'Fallon's paper on the Children'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.

Inventory size analysis

Above we summarized prior research suggesting that the SCT worked well with shorter versions. Here we evaluate inventory length variation with the STAGES assessment particularly.

A shorter, 16-stem, inventory. Given the strong results on the reliability of the STAGES SCT mentioned above, we decided to experiment with a shorter inventory containing 16 stems. This gives us an inventory that subjects can take in half the time, which is important for some applications.

We selected 16 stems from the General inventory which we had previously gotten a high Cronbach's alpha on. We chose sentence starters that would point to each of the four quadrants. The stems were: 1, 2, 5, 6, 8, 18, 19, 20, 21, 23, 25, 27, 28, 30, 34, 35 (from Table 1). The Cronbach's alpha for the 16-item inventory (0.98) is in the "excellent" range.

Cronbach's-Mesbah analysis. The paper "Investigating the Validity of the Ogive Aggregation Method, Including the use of Rasch Analysis, for the Sentence Completion Test and the STAGES Model" (Murray, this issue) includes additional analysis related to test length, which we summarize next. Figure 2 shows a Cronbachs-Mesbah curve illustrating the Cronbachs-alpha score for successively fewer test stems (with the worst correlated items removed first). It 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).

From a purely statistical perspective, the SCT doesn't need so many items (assuming this pattern was also true of the WUSCT or MAP variations of the SCT – though we don't have item-level data for those). However from a qualitative research agenda, and for coaching and debrief purposes, a full set of 36 responses is important for identifying themes and patterns in the detailed reports and coaching sessions given for individualized SCT assessments. Thus for the

purposes of sharing qualitative information, and not just a final statistical score, the longer inventory is preferable.

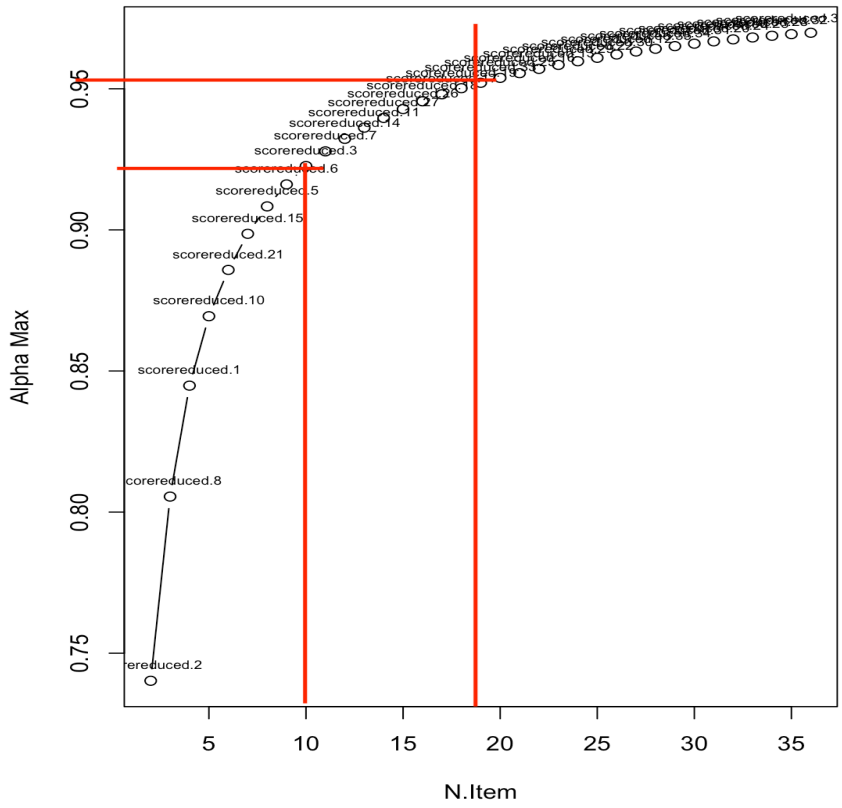


Figure 2. Cronbach-Mesbah curve of alpha vs. number of items.

Conclusions

Over the past 4 years we have experimented with the efficient and effective adaptability of the STAGES SCT Inventory. Beginning with the alteration of the General Inventory, we learned that we could change up to 9 stems all at once in a 36 item inventory, and still have very high internal consistency. This prompted us to experiment with other uses of this process and the idea of the specialty inventories was born.

The resultant findings indicate that the STAGES model allows for efficient introduction and experimentation with variations on 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, and climate change, psychological reflection, and a children's version of the SCT. Most of the specialty protocols replace 6 of the General protocol stems with theme-specific stems. However generally we have an indication that we can change up to 9 stems as depicted by our research on the new General inventory. We are now experimenting with variations on this theme.

We have shown results for internal consistency studies of (1) the STAGES General Protocol, (2) Six specialty protocols, and (3) a shorter 16-item inventory. We have shown that for all of these inventories there is excellent internal consistency (excluding the special case of the Children's study).

Our research not only gives evidence about the reliability of the particular specialty inventories, but also shows evidence that the SCT is quite robust to the addition of new stems, assuming they are well-written.

Our research 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.

These statistical results give us an underlying confidence that these variations on the regular inventory are internally consistent.

Once we have this assurance we then do the qualitative research that allows us to extract meaning from what the test takers actually say. For example, we have done qualitative research on the Love inventory. We selected all the Love stem scores at each developmental level and created categories out of their responses. The resultant understanding points to the perspectives people take on Love at each developmental level. As well we can see the evolutionary trajectory on the test takers views on love. The addition of the qualitative research can support curriculum development, policy development, planning, and particular interests of various audiences.

Future Research

Based on the research on adapting the General inventory in length and specialty there is further research we aspire to.

1. We are currently working with domain experts to design additional specialty protocols in these areas: relationships, beliefs related to a specific religious tradition, spirituality, money, hope, dementia, ethics, and parenting.

2. Since we can see that there is a variation from 5 to 9 stem changes and still maintain internal consistency, we will be experimenting with variations of these numbers for other specialty inventories.

3. Since the 5-9 stems isolated from the general inventory still have a "good" internal consistency we will look into using only the specialty stems to indicate the developmental perspectives that people take on the specialty only. This would involve much shorter inventories for people to take which might bring a research advantage, if there is enough material to do qualitative research.

4. Once the quantitative research is complete, qualitative research yields more useful and practical day to day value for our clients. Thus we will be continuing to do serious qualitative research on all specialty inventories and to perfect the relationship between the

quantitative and qualitative interface for the benefits of our services to others. As well, we will look into how to develop curriculum, planning, policy, and service to each specialty area and the audiences that resonate with them.

References

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

Brier, S. (2013). Cybersemiotics: A New Foundation for Transdisciplinary Theory of Information, Cognition, Meaningful Communication and the Interaction Between Nature and Culture. *Integral Review: A Transdisciplinary & Transcultural Journal for New Thought, Research, & Praxis, 9*(2).

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.

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

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

Fischer, K. W., Hand, H. H., & Russell, S. (1984). The development of abstractions in adolescence and adulthood. *Beyond formal operations, 1*, 43-73.

George, D., & Mallery, P. (2003). *SPSS for Windows step by step: A simple guide and reference. 11.0 update* (4th ed.). Boston: Allyn & Bacon.

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

Loevinger, J. (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*. London: Psychology Press.

Miniard, A, C. (2009). *Construction of a Scoring Manual for the Sentence Stem A Good Boss – For the Sentence Completion Test Integral* (SCTi-MAP) ETD Archive. 491. https://engagedscholarship.csuohio.edu/etdarchive/491

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

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. & O'Fallon, T. (2020). Summary of STAGES validation research. *Integral Review*. This issue.

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. J. (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.

Sokal, R. R., & Rohlf, F. J. (1962). The comparison of dendrograms by objective methods. *Taxon, 11*(2), 33-40.

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

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

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