The validation of a new scoring method for assessing ego development based on three dimensions of language

Terri O’Fallon a,*, Nayak Polissar b, Moni Blazej Neradilek b, Tom Murray c

a Developmental Research Institute, STAGES International, 608 NE Washington Street, Lewistown MT, USA
b The Mountain-Whisper-Light Statistics, 1827 23rd Ave. East, Seattle, WA 98112, USA
c Open Way Solutions LLC, 36 Fruit Street, Northampton MA, 01060, USA

ARTICLE INFO

Keywords:
Psychology
Meaning-making complexity
Ego development
Construct-developmental theory
Scoring methods

ABSTRACT

We describe research on the validity of a new theoretical framework and scoring methodology, called STAGES, for sentence completion tests of meaning-making maturity or complexity (also called ego development or perspective-taking capacity). STAGES builds upon research on the substantially validated Washington University Sentence Completion Test of Jane Loevinger as updated by Susanne Cook-Greuter. STAGES proposes an underlying structural explanation for the Cook-Greuter system based on three dimensions. Two of these are polar factors: individual/collective, and passive/active; and the third is a categorization of the sophistication of the types of objects referred to (i.e. as concrete, subtle/abstract, or “metaware”). We describe two validation studies for the STAGES scoring method and model. The first is a replication study of concurrent validity, using 73 inventories to test the hypothesis that the STAGES scoring method replicates the Cook-Greuter scoring method. Using the weighted Kappa statistic, we demonstrate a very strong match between the two methods, confirming the first hypothesis. This study includes levels up to and including Strategist (i.e. a substantial percentage of test-takers from most populations). Levels above Strategist were validated using another method because (1) there is less Cook-Greuter data available at these levels, and (2) the two scoring methods diverge sufficiently, making comparison difficult. The second study, of 71 inventories, attempts to validate the STAGES scoring method at levels above Strategist by testing the inter-rater reliability among four scorers. The inter-rater reliability above Strategist, using the weighted Kappa statistic, was found to be moderate to substantial, indicating that the instrument and scoring method has internal validity for these four, rare higher levels. Additionally, the inter-rater reliability over all STAGES levels were found to be very strong.

1. Introduction

Among scholars and pundits who analyze global trajectories in human capacities, there is increasing calls for two types of skills: the so-called “soft” skills of social/emotional intelligence and self-knowledge and the complex higher-order thinking skills for responding to “volatility, uncertainty, complexity and ambiguity” (VUCA, or “wicked problems”) (see McChrystal et al., 2015; Conklin, 2005). Though these skill sets can be understood separately, they are also closely related, primarily because the interpersonal skills required in the workplace and the social mastery involve significant complexity and uncertainty in the social domain. For instance, literature on 21st Century education and workforce development calls for self-reflective and critical thinking skills, communication and empathy skills, multi-stakeholder perspective consideration, robustness within paradox and uncertainty, and understanding of systems and networks (NSTA, 2011; Clark et al., 2009; Scardamalia et al., 2012), while a similar set of skills has been suggested to be a requisite for robust citizen participation in democracies (Muhlberger and Weber, 2006; Rosenberg, 2007; Murray, 2017).

Assessing these human capacities is a crucial aspect of supporting their growth in individuals and in society as a whole. Valid assessments should be derived from sound psychological theories. There are numerous theories and frameworks addressing the large set of capacities mentioned above. Our work centers on “construct-developmental” theories of human meaning-making, self-understanding, and perspective-taking. The contemporary understanding of human psychology acknowledges that adults can psychologically and cognitively change and grow over their lifespan, not only in terms of storing new memories,
learning new information and skills, and acquiring new knowledge but also in growing *developmentally* to change one’s most basic understandings of the self, the other, and the world. These holistic theories frame psychological maturity and human potential in terms of the complexity of one’s worldview and meaning-making about the previously mentioned three domains (and, importantly, relationships among these three domains). This field of “adult development” includes research on several closely related constructs including ego development, meaning-making sophistication, perspective-taking complexity, and wisdom skills (Fischer, 1980; Hall, 1994; Wigginsworth, 2012; Hy and Loewinger, 1989; Loewinger and Wessler, 1970). This paper explores a new theoretical model and assessment method for such capacities called STAGES. In this paper, we will (1) describe a theoretical model that proposes a small set of underlying factors that drive the developmental growth described by other models and (2) evaluate the validity of a new scoring system based on this model.

1.1. Background on construct-developmental theories

Early works in the developmental theory lineage include those of James Mark Baldwin (1901), and Jean Piaget (1969), from which many other developmental models emerged (e.g. moral development: see Gilligan, 1993; Kohlberg, 1973; and values development: see Graves, 2002; Hall, 1994). Presently, developmental scales are commonly used in psychology, counseling, child development, leadership, and other areas (Forman, 2010; Torbert and Livne-Tarandach, 2009; Wilber, 2000).

Although many of these research projects have focused on narrowly defined skills, some theories have successfully established the validity of more overarching constructs. Two of these are noteworthy: Kegan’s construct-developmental model (1994) and Loewinger’s ego development model (Hy and Loewinger, 1989; Loewinger and Wessler, 1970) chart very similar (conceptually correlated) territory in the evolution of psychological/cognitive “meaning-making” in terms of a hierarchical sequence of stages. These psychological frameworks have been empirically derived and validated. Our work on developmental assessment extends Loewinger’s research lineage.

Loewinger’s model of ego development was intricately linked to her assessment instrument, the Washington University Sentence Completion test (WUSC) (Hy and Loewinger, 1989; Loewinger and Wessler, 1970). This assessment, later updated by Cook-Greuter (1999), is hereafter referred to as the Loewinger/Cook-Greuter model or simply CG/L. The CG/L test differs from related instruments that use self-rating or dilemma-solving activities because it is a “proactive” test in which subjects complete sentence starters, responding freely without a need to produce a “correct” or superior answer. Browning (1987, p. 113) notes that ego development theorists “[postulate] 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.” When we refer to “development” in this paper, unless specified otherwise, we mean ego development or, equivalently, meaning-making maturity, perspective-taking level, or development of higher-level cognition and awareness.

The WUSC test is one of the most researched developmental scales used in psychology today. The literature on Loewinger’s ego development model is quite extensive and includes over 40 years of meta-analyses and critical overviews, substantially supporting its validity and usefulness (Cohn and Westenberg, 2004; Manners and Durkin, 2001; Holt, 1980; Novy and Francis, 1992; Jespersen et al., 2013; Westenberg et al., 2004; Forman, 2010). According to an overview by Westenberg et al. (2004), the WUSC test has quite 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 more studies have followed since 2004 (Torbert and Livne-Tarandach, 2009).

Cook-Greuter (1999) advanced the original scoring system by adding a structural logic to Loewinger’s theory, strengthening it from a “soft” construct to a “hard” construct, by linking the person perspectives to the stages (p. 77). This provided a coherent ego theory that could support the developmental trajectory (p. 72–76), as a trajectory pattern for the developmental structure of the ego developmental scale which was missing until that point.

She also verified a new later stage (Construct Aware) and proposed a further stage called Unitive. She streamlined the scoring process, but only for the two new stages, by creating scoring rules for these two levels that apply to all stems. For all of the previous seven levels, she continued to use the WUSC method, which has a different set of exemplars for each stem of each level (plus some general rules intended to cover the rare cases when an exemplar-match cannot be found).

1.2. Objectives for the STAGES research

The STAGES model and assessment was formulated to build upon the Cook-Greuter-Loewinger ego development framework. STAGES retains the valuable base of its predecessors with the addition of five objectives to update and strengthen the ego development model and its assessment:

1. Changing the scoring system from stem- and exemplar-based to generic and heuristic-based.
2. Incorporating person perspectives more completely into the scoring system.
3. Developing definitions of person perspective (and thus of developmental levels) that are independent on specific content and word meanings (i.e. moving from a content-based to a structure-based assessment of language).
4. Including a relatively consistent “step” or “width” in the progression of developmental stages.
5. Supporting a deeper understanding and more specific definition of the highest stages of development.

The first objective involves changing the scoring from an exemplar-matching approach to a general set of scoring heuristics that apply to all stages and all sentence completions. The standard Loewinger and Cook-Greuter sentence completion projective test has 36 sentence starters (“stems”), such as “Raising a family...” which the test taker completes (e.g. “...is a joy”). Sentence completions vary from a few words to full paragraphs, and sometimes multiple paragraphs. Other versions of the Loewinger WUSC have used as few as 18 sentence starters to more than 36; however, current models use 30–36 sentence starters (Cook-Greuter, 1999; Torbert and Livne-Tarandach, 2009). The completed set of sentences from an individual is referred to as an “inventory.” Stems are chosen to address a holistic set of life themes (self, relationship, society, work, family, etc.) that, in a sense, triangulate the measurement of one overarching construct (i.e. “ego development”) from many perspectives. The 36 scores in an inventory are combined into a total developmental score (“TPR,” Total Protocol Rating) for the inventory (see the Appendix for a description of the cutoff method developed by Loewinger).

A scorer using the Loewinger and CG/L system consults a scoring manual that comprises thousands of example sentence completions organized by stem, stage, and theme. The scorer attempts to match a sentence completion with an example or a thematic example category. If no match can be found, more general heuristics defined for each level (vs. for each level and stem) are used. For the two highest levels, Cook-

---

1 Soft stage theories are descriptive narratives that attempt to explain the observed phenomena, such as how meaning-making changes with age and type of experience. Hard stage theories make stronger claims about the underlying causal drivers of development.
Greuter's system relies on heuristics in addition to exemplars. A current scoring manual contains 16,000 + examples, an average of about 50 for each of the nine levels for each of the 36 stems; organized into approximately 10–12 thematic categories for each of the 324 (36^2) stem-and-level sections of the manual.

Though the example-matching method was chosen by Loevinger for specific reasons, it has certain drawbacks. One issue is that matching to exemplars can be tedious and time-consuming. (Though highly skilled scorers, with years of experience, have memorized the gist of most categories and can score most inventories without consulting the manual.) It is also time-consuming to add new levels or sentence starters in such a system, making it less agile and adaptable—it requires the collection and validation of excessive amounts of data to define exemplars and heuristic rules (e.g. see Miniard, 2009). Thus our first objective was to recast the entire scoring system in terms of one set of principles that can be applied to any sentence starter and all developmental levels. As will be described, the new scoring system is based on evaluating three dimensions or parameters of language, corresponding to the theoretical model's three "drivers" of development.

The second objective involves incorporating person perspectives more completely into the scoring system. In her update to the Loevinger system, Cook-Greuter attempted to address her notion that Loevinger's model was "lacking an underlying structural logic" (Cook-Greuter, 1999, p. 76). She corrected this putative lack by tying the developmental stages to a sequence of "person perspectives" (e.g., first-person perspective, second-person perspective, third-person perspective, etc.,...sometimes referred to as worldviews) to add a "structural logic." This transformed the model from a "soft stage theory" toward more of a "hard stage theory;" (p. 76, Cook-Greuter, 1999). The articulation of person perspectives in Cook-Greuter's model served an explanatory and descriptive function at the theoretical level but was not well integrated into the exemplar-based scoring procedure. Our second goal was to fully integrate and extend the person-perspective framework within the model and the scoring system.

A third, related, objective involves generating specific, enduring, and fundamental definitions for the person perspectives that represent each stage that would 1) serve as underlying mechanisms for driving development and 2) capture the developmental trajectory of cognition and awareness. This would firmly move the theory from a content-based to a structure-based foundation, i.e. the scoring method and theoretical model would no longer depend on the meanings of specific words or concepts but would depend on the more complex structural properties of language and reason. Word meaning changes through time, and exemplar-based systems risk losing relevance as culture changes (or they require a painstaking process of semantic re-calibration).

The fourth objective was to have a relatively consistent "step" or "width" in the progression of developmental stages. This would integrate an important property of contemporary Neo-Piagetian models within the original Loevinger framework. The number and demarcations of Loevinger and Cook-Greuter's levels have evolved somewhat haphazardly, based on practical considerations, and do not seem to indicate that each level represents a more-or-less equal "distance" along the developmental trajectory.

In parallel with research on construct-developmental theories of meaning-making (Loevinger and Kegan) is research on developmental theories in the "Neo-Piagetian" tradition that proposed domain-independent underlying mechanisms for the development of any skill or capacity (vs. Loevinger's model of a single, though widely holistic, capacity). The most advanced model known of these are Kurt Fisher's Skill Theory (Fischer, 1980; Fischer and Zheng, 2002) and Michael Commons Hierarchical Complexity Theory (HCT, Commons, 2008; Commons et al., 1998). Fisher and Commons independently proposed and validated surprisingly similar developmental models (which were later integrated through Dawson's framework, Dawson, 2004). Similar to Loevinger's framework, they describe development in terms of an invariant sequence of levels, but unlike Loevinger who provides empirically derived descriptions of each level, the Neo-Piagetian frameworks propose underlying mechanisms driving development. These mechanisms describe how a level coordinates and transforms the skills or capacities of the prior level. Skill Theory and HCT are designed to assess relatively narrow or specific skills or "lines" of development, while meaning-making development is a more extensive holistic capacity not easily captured by these Neo-Piagetian theories. Therefore, one of our goals was to integrate the principles of the Neo-Piagetian models with the relevant work in the construct-developmental tradition of Loevinger and Kegan.

Objective five involves acquiring a deeper understanding and more specific definition of the highest stages of development. Preliminary research had indicated that more structure, definition, and clarity could be added to the top two levels of Cook-Greuter's model. Our group had more data on these higher levels, and we inferred that this territory could be explained better as a "third tier" containing four levels (as discussed later).

We describe our new model and report a successful validation study in the following sections.

2. The STAGES model and scoring methodology

2.1. The STAGES model overview

The STAGES model proposes that the levels of the CG/L developmental model can be explained and defined in terms of a small set of underlying properties (or "parameters"), which constitute the definition of each person perspective. Specifically, the developmental level of sentence completion (or any text) can be determined by answering three questions that address three parameters or dimensions: (1) What is the Tier (i.e. category of object awareness)—Concrete, Subtle, or MetAware? This marks the trajectory of one's ability to understand (conceive of) objects of different levels of complexity, abstraction, and/or nuance. (2) Does it foreground Individual or Collective objects? This highlights whether the experience is all about "me" or about "we" (including relationships, groups, or systems as described below). Finally, (3) is the cognitive orientation receptive (simple passive), active (simple active), reciprocal (complex with passive predominating), or interpenetrative (complex with active predominating)? This question marks the developmental progression of increasing levels of complexity within the tier structure oriented to a particular type of object. The first parameter (dimension) has three values (three tiers) and the second and third have two values (Individual vs. Collective and Active vs. Passive); therefore, there are 12 possible outcomes (3 × 2 × 2), and thus, there are 12 levels in the STAGES model. These are illustrated in Figure 1.

These 12 levels correspond to the nine levels of Cook-Greuter's model if three of the Cook-Greuter levels were refined through sub-division into two categories (i.e. some of the Cook-Greuter levels merge a passive and an active part—see the asterisks in Table 1). For instance, a sentence about Subtle Individual objects, Passively oriented, is scored 3.0, while text focusing primarily on a Concrete Collective object, Actively oriented, is scored 2.5.

Table 1 shows another visualization of these three scoring questions or parameters. The "common name" in the table provides each numerically labeled stage a descriptive handle—it is not used to define the level in the scoring procedure. The sequence of stage numbers represents a sequence of person perspectives or worldviews.

As mentioned previously, one of Cook-Greuter's innovations to Loevinger's model was to start mapping "person perspectives" onto the developmental sequence. Literature in psychology and philosophy elaborates on the nature, function, and development of assuming a first, second, or third-person perspective with respect to an object or event (e.g. Selman, 1971; Habermas, 1990). The second-person perspective involves stepping outside of the self (first person) to imagine how another individual might perceive/cognize or interpret something; while the third-person perspective involves stepping back even further to observe how the generic or prototypical rational human being would perceive or
interpret something. This system has been extended into fourth, fifth, and sixth-person perspectives in which each perspective involves being able to observe the prior person perspective as an object. In the ego development models, it is most suitable to interpret this sequence casually as each person perspective in ego development is associated with, but not completely defined by, the concept of person perspective.

Stage 1.0 represents the early first-person perspective; 1.5 represents the late first-person perspective; 2.0 represents the early second-person perspective, etc. Therefore, the “person perspective” framework of Cook-Greuter has been carried forward and extended, first by categorizing early (passive, “0”) vs. late (active, “5”) phases of each person perspective; and second, by extending the scheme up to the sixth-person perspective, and third, by expanding the scale as required to assign two levels to all person perspectives.

As STAGES was derived from the CG/L model and aims to enhance and extend it, the common names have mostly been mostly adopted from the CG/L model. The “other names” in the table have been provide to help readers familiarize with related models, including those of Cook-Greuter and Torbert, to coordinate the different names given to these levels.

### 2.2. Scoring questions and an example of scoring

Next we provide an example of the scoring method. The first scoring question about the tier (see Figure 1) specifies the general type of object one is aware of. By object we are referring generally to anything one can focus one’s awareness on and refer to, including physical things, subjective experiences, processes, properties, abstract ideas, etc. Concrete stages apprehend concrete objects of which there are two types. First are phenomena that are perceivable through a direct experience of the exterior senses. Examples include cars, a church, or rules in sports and card games. Second are those same phenomena that one can experience through their interior senses (visualization, interior hearing, and interior feelings or emotions). Subtle stages apprehend more abstract objects, including phenomena that one cannot form distinct and accurate images of, or hear sounds about, or touch as they do with their exterior or interior senses. Examples include brainstorming, reasoning, contexts, complex adaptive systems, models, values, determinism, democracy, and square roots. Entrance into the Subtle tier corresponds roughly to the transition from Piaget’s Concrete Operational to Formal Operational thinking and includes abilities in the arena of abstract, logical, and systematic reasoning.

MetAware stages apprehend even more subtle objects such as the capacity to examine one’s awareness of concrete and subtle objects to clearly determine the previously assumed constructions of the mind such as word meaning, boundaries, and the reification of time and space. MetAware “objects” are more similar to processes or properties that are perceived to permeate, pervade, or underlie reality or experience. Examples include what has been called witnessing of consciousness or experiencing ideas or identity formation in the mind as it happens (and thus experiencing the emptiness aspect of the self or the meaning-making processes). Fullness is a characteristic as well as emptiness, for instance, experiencing a sense of oneness, life energy, or beauty pervading everything.

The second scoring question is about whether the primary object being mentioned is an individual or collective object. Collective objects are relationships, groups, processes, or systems involving two to many individual objects. It is relatively straightforward to describe collective objects in the concrete tier—e.g., flocks, teams, towns, families, etc. (including religions and nations, when experienced in a concrete way). The Concrete Collective level also involves early (concrete) forms of relationally, care, and perspective-taking (e.g. being able to imagine that hitting another person hurts them). Subtle collectives are systems and

---

### Table 1. STAGES Tiers & Repeating Principles Asterisks (*) show levels added in STAGES vs. CG/L model.

| STAGES Levels | Common Name | Other Names |
|---------------|-------------|-------------|
| **CONCRETE TIER** | | |
| 1.0 Concrete Individual Receptive | Impulsive | |
| 1.5 Concrete Individual Active | Egocentric | Opportunistic |
| 2.0 Concrete Collective Reciprocal | Rule oriented* (Delta/3) | |
| 2.5 Concrete Collective Interpenetrative | Conformist | Diplomat |
| **SUBTLE TIER** | | |
| 3.0 Subtle Individual Receptive | Expert | |
| 3.5 Subtle Individual Active | Achiever | Conscientious |
| 4.0 Subtle Collective Reciprocal | Pluralist | Individualist |
| 4.5 Subtle Collective Interpenetrative | Strategist | Autonomous |
| **METWARE TIER** | | |
| 5.0 MetAware Individual Receptive | Construct Aware | Alchemist |
| 5.5 MetAware Individual Active | Transpersonal* (Unitive?) | |
| 6.0 MetAware Collective Reciprocal | Universal | |
| 6.5 MetAware Collective Interpenetrative | Illumined* | |

---

Note that by apprehending an object (or property, process) we mean a type of experience of it and not a mere intellectual idea of it. One might be able to talk about a “cultural narrative” or witnessing awareness as an idea one has learned about in a classroom but to really apprehend, it means one can notice it and reflect on its existence in actual lived contexts.
interrelationships of subtle or abstract things. Examples include apperception of value systems, cultural narratives, ecosystems, family system dynamics, situational contexts, as well as projections, introjections, and complex holistic world-systems.

A MetAware collective is one whole that includes, subsumes or transcends, concrete and subtle manifestations. It is perceived as that which permeates or underlies all experience. Such apperceptions can include a sense of emptiness (vanishing) and/or fullness (omnipresence) of the timeless, boundless, and beingness. However, though these constructs might sound esoteric or “New Age,” they are intended to describe the verbal behavior of actual individuals in these later stages.

The third scoring question is about the Passive/Active dimension or, equivalently, determines whether the text is Receptive or Active (for Individual perspectives) or Reciprocal or Interpenetrative (for Collective perspectives). Grammar and sentence structure are used to determine scoring for the third question. Receptive sentence completion tends to use passive language, an active one tends to use active language, a reciprocal one tends to use passive and active language with passive language prevailing, with an interpenetrative one tends to mix/integrate active and passive with an emphasis on active language. Active orientation is also indicated by ownership (“my,” “our”) language. These grammar and sentence clues, augmented by the meaning in the sentence completions, help the scorer derive a final stage for each of the 36 completions.

It is beyond our scope to provide a full description of each stage or the insights about human development and meaning-making that the STAGES model supports (see descriptions of STAGES in O’Fallon, 2011, 2013; Murray, 2017; Integral Review, 2020 in process).

Applying the Scoring Questions to an Example Sentence. The scoring rules follow scoring principles derived from the three primary STAGES dimensions (with advanced scoring using the fourth dimension, Interior/Exterior, for sub-level determination, which is not discussed in this paper). All text is scored using these same scoring rules. For instance,

Sentence starter: “A good child_”; completion: “is a friend.”

Based on question 1, “Is the response Concrete, Subtle or MetAware?”
We determine that this is a completion in the concrete tier because in this context a friend is a concrete person. This eliminates all the stages in the Subtle and MetAware tiers and narrows the choices down to the four stages in the concrete tier.

Based on question 2, “Is the response individual (it’s all about me) or collective (it’s about a we, us, or system)?” We can see that this is about two people, me and a friend, i.e. a relationship. So, a collective score is required rather than an individual one. This eliminates all the stages that have an individual orientation. This leaves two collective stages to choose from (2.0 and 2.5).

Based on question 3, “Is the response receptive, active, reciprocal or interpenetrative?” (We usually use this wording instead of “Is it Passive or Active?”), because, for example “reciprocal” better captures the meaning of Passive-Collective.) As a collective completion, our choices are either reciprocal or interpenetrative. The verb “is” is passive therefore the best choice is the reciprocal quality and we have the final textual scoring of Concrete, Collective, Reciprocal, i.e. 2.0 early second-person perspective or “Rule Oriented” as shown in Table 1.

In practice, learning the nuances of scoring is more complex than indicated in the example above—one trains for about a year to become a certified scorer for the Sentence Completion Test. Nevertheless, the general principles behind the STAGES parameters can be presented in short workshops to guide an overall understanding of perspective-taking and worldview orientation. Examples of scoring at various stages are provided in Appendix 2.

Combining the 36 sentence completions scores into a final score.
Once the 36 sentences of an inventory have been scored using the new scoring system, they are combined into a final score using the “cutoff” values developed by Loevinger and continued with Cook-Greuter. A cutoff is the required number of responses out of the 36 that one must score at or above any level to have a final score at that level. Appendix 1 includes a table showing the cutoff values used in the STAGES method based on the CG/L cutoff values. This method is used because a simple mean (or mode) does not capture the intuitive understanding that the evidence of later level sentences should have more weight in the total “center of gravity” score.

2.3. Adding stages to the CG/L system, and defining MetAware stages

In Table 1 the asterisks indicate levels added to CG/L in the STAGES model, and the “Other Names” column helps coordinate between the CG/L and STAGES levels in case level names differ. Below we describe the differences between the two frameworks.

Early levels. Using the STAGES model lens, it was apparent that the CG/L stage of “Diplomat” was a whole person perspective composed of both early and late phases. Therefore, this stage was separated into two, 2.0 and 2.5, in the STAGES model. Doing so actually revived a stage (“Delta”) that existed in the earlier versions of Loevinger’s model that was combined with “Diplomat” in more recent versions of the Loevinger and CG models.

Late levels. Cook-Greuter’s research (1999) extended the Loevinger stages to include two higher levels of development which she labeled Construct Aware and Unitive. Unitive is a holding category for everything that is above Construct Aware. O’Fallon’s analysis showed that the data available at levels above Strategist fit appropriately into the STAGES system which further divides the fifth- and sixth-person perspectives into the early or Passive and later or Active phases of each person perspective, resulting in definitions of the four stages in the MetAware tier: 5.0, 5.5, 6.0, and 6.5.

However, the level definitions in the two systems do not align as well in the MetAware Tier as they do in the lower tiers. The MetAware stages (those above Strategist in Cook-Greuter’s model) are the least comprehended (in all models) and researchers have the least data about them. Their definitions, in both models, are thus the most tentative as well as diverge the most between models. CG/L includes two stages above Strategies (4.5) while STAGES has four.

The STAGES 5.0 level is primarily the same as the CG/L Construct Aware stage, and the data from CG/L Unitive stage is distributed primarily into 5.5, and 6.0, (however some inventories at 4.5 (Strategist) in the CG/L model have definitions corresponding to 5.5 in STAGES). As the two models differ significantly above Strategist, this study uses a separate statistical method for the MetAware Tier. For levels up to Strategist (i.e. tiers 1 and 2), we conducted a rigorous “replication study” of the concurrent validity of the system in comparison with the CG/L system. For higher levels (MetAware or Tier 3), we conducted an inter-rater reliability assessment.

To summarize, when defining the person perspective parameters, it seemed critical to have a measuring stick representing both an early and a late perspective in each stage. Missing in the CG/L system was the early second-person perspective which we retrieved from earlier versions of the Loevinger scale and the late fifth-person perspective and late sixth-person perspective which were not distinctly represented in the Cook-

An alternate “TWS” (total weighted score) method using a weighted sum is sometimes used which yields an integer value usually between 300 and 600. But the Ogive (cutoff) method maintains the notion, preferred by Loevinger, that the scale should return categorical (ordinal) rather than the numerical results. Note that we are developing a new method for calculating the STAGES total protocol score but it is not yet finalized, and at any rate, we use the Ogive method here for comparison with prior scoring systems.

5

---

3 The fact that these stages are “later” is evidenced by longitudinal studies that we are in the process of reporting for publication.
Greuter update. These missing perspectives were intentionally added to the STAGES scoring system to provide an even representation between and across the perspectival stages. That STAGES is based on an underlying structure of repeating patterns (parameters) allows us to predict the nature of ever higher stages, where less data is available—Loevinger and Cook-Greuter's methods are not designed to speculate about stages lacking significant data. Of course, all theory-based speculation must be corroborated empirically.

3. Method — Evaluating the STAGES scoring system

Now, we describe our statistical validation of the STAGES model, which includes a replication study comparing STAGES scoring to CG/L scoring for Tiers 1-2 and an inter-rater reliability (IRR) analysis of Tier 3 (MetAware) and of all levels combined.

The Appendix “Development of the STAGES Model and Scoring Rules” describes the “grounded theory” approach used to construct the model and the assessment. This section describes the empirical validation of the assessment.

Till date, about 10 individuals have been certified to score using the STAGES model and more are under supervision for certification. The first cohort of four trained scorers participated in the validity study described later in this paper. A scoring trainee must score approximately 100 inventories to learn to score accurately under supervision with feedback. To be certified to score, they must achieve an 85% inventory-level agreement on the final stage score compared to a master scorer. All stages are represented equally within the set of practice inventories.5

3.1. Method overview

Beginning with a set of approximately 750 inventories, most of which were scored previously using the Cook-Greuter (CG/L) method, 142 were selected for this study using sampling methods described later. 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). The goal was to demonstrate concurrent validity of the scoring method through a replicability study between the two methods and also to demonstrate consistency of the measurement through an inter-rater reliability method.6 Additional validity metrics are summarized later, though not detailed in this paper.

Because the two systems have relatively different definitions of levels above 4.5, the replicability study was conducted for stages up to and including 4.5. For the purpose of this report, we will call this the Tier 1-2 data set, and the data for levels higher than 4.5 (i.e., above Strategist) will be called the Tier 3 data set (note that this description uses STAGES terminology to classify based on the prior CG/L scoring of the data—e.g., the CG/L model does not mention tiers). Because of the expected divergence in level definitions in Tier 3, for that tier, an IRR (replicability) study was conducted as an indication of test consistency. 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 and Livne-Tarandach, 2009). About half of the 142 inventories were used for the replication study of Tiers 1 and 2 and half for the IRR study of Tier 3.

3.2. Data sources

Pacific Integral6 (PI), an organization whose activities include developmental assessments and the development of educational and social change technologies, had used the Loevinger scale—as updated by Cook-Greuter—for about 10 years before switching to the STAGES model. The entire data available for the project included the following:

- At the beginning of this research project, about 750 inventories were targeted from PI's database, gathered from participants who had taken the inventory in PI's “Generating Transformative Change” program along with inventories from individuals and organizations who requested testing through PI. This data had previously been scored using the Cook-Greuter (CG) method. It had been scored by four different individuals, including O'Fallon and Cook-Greuter. Data was randomly sampled from this set using a stratified sampling method as described below.
  - The STAGES model splits the CG/L “Diplomat” level into two levels: 2.0 and 2.5. Considering our comparison of the two systems' scores for these levels, the following aspects are noteworthy.
  - First, in the comparison of the two systems, the inventories’ scores as Diplomat in CG/L could be, in terms of STAGES, at 2.0 or 2.5. CG/L Diplomat was mapped to STAGES 2.5 (as opposed to 2.0) because the definition of Diplomat is conceptually more similar to the STAGES definition of 2.5 than its 2.0. This conflation of levels (that some Diplomat 2.5's are “actually” 2.0) would serve only to worsen the statistical (Kappa) comparison. That is, this mismatch does not worsen the validity or magnitude of the research results.
  - Second, though the standard CG/L system does not have a 2.0 equivalent score, in the past, it had a level called “D3” (Delta) that was excluded from the model (i.e. combined with Diplomat) because there was insufficient evidence or theoretical reason to differentiate it from Diplomat within the exemplar data pool. However, archived data did exist from scoring that included D3. An additional set of six archived inventories were obtained from Cook-Greuter directly that had been scored by her at the “D3” (Delta) level.
  - Third, in the statistical results, we will present comparison statistics for both 2.0 and 2.5 separated and 2.0 and 2.5 combined.
  - As the set of data previously scored using the CG method had very few inventories scored above Construct Aware (5.0), the IRR study of Tier 3 includes 55 additional later level inventories that were not scorable by the CG model rules of the two later levels in her research. They were included in the data set and scored for the first time in the STAGES research study.

All data was in the form of frequency distributions, i.e., the number of sentence completions rated at each level for each inventory.5

Tier 1-2 Study, n=73. The data available for the Tier 1-2 analysis of replicability and IRR included inventories rated at Strategist (4.5) and below from the PI database in addition to the six D3 inventories studied in the replicability study of Tiers 1 and 2.

---

5 The training inventories have been previously scored and agreed upon by master scorers. “Master scorer” is an informal term we use for certified scorers with exceptional experience. The trainees score the inventories with partners and check their work against each other and against the master scorers, receiving feedback from several sources as they learn, in a training and certification process that typically takes approximately one year.

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

7 PI is an organizational affiliation of one of the authors (TO). Since this study PI has split into two companies, Pacific Integral and STAGES International with the latter continuing the practice and research of developmental scoring.

---

8 This data format was the typical record keeping method used in the Loevinger tradition before electronic methods were available. This meant that for some of the data, the scores for each of the 36 sentences was not available but only the overall frequency distribution. This prohibited certain common forms of psychometric test analysis, such as internal consistency, usually measured with Cronbach's Alpha.

9 The total of 144 inventories added together for both studies (71 + 73) is greater than 142 because two inventories qualified to be included in both studies.
mentioned previously. A stratified sampling method, described later, was used to select all the 73 inventories used in the Tier 1-2 study.

**Tier 3 Study, n=71.** As mentioned previously, as the two systems have relatively different definitions of levels in Tier 3 (5.0, 5.5, 6.0, 6.5) and there were very few CG/L-scored inventories above 5.0 (Construct Aware), an IRR analysis was performed with the Tier 3 data. The Tier 3 analysis used 16 inventories from the original CG-scored database and the 55 “additional” inventories mentioned previously for a total of 71 inventories.

**All-Tier Results, n=142** [10]: We also present an IRR analysis for the full data set of 142 inventories, combining Tiers 1–2 and 3.

The demographic characteristics of the 142 participants are as follows. The age spanned from 19 to 69 years, averaging about 40. About 45% of the subjects were female (vs. male). Among those specifying education levels (about 80%), 4% were at doctoral level, 40% at master’s level (or equivalent), 39% at bachelor’s level, and 18% at high school (or not finishing high school) level. Subjects were from a variety of locations around the world (Ethiopia, Germany, France, The UK, Ireland, United States, Australia, New Zealand, Canada, Russia, Kosovo, Pakistan, China, Hungary) All the participants spoke English as their first or second language. As well, a variety of professions are represented: (student, lawyers, consultants, psychotherapists, spiritual teachers, high school and university teachers, wanderer, organic farmer, construction workers, CEO’s, doctors, project directors, government workers, researchers, coaches, people who weren’t working due to disabilities, IT data coders). Some were chosen from a Cook-Greuter database which included prison populations and populations from Mental Health institutions. As has been observed in many prior studies in the Loehvinger tradition, average developmental level roughly increased with both age and education.

**3.3. Data sampling and scorer selection**

The method for selecting data for the Tier-3 IRR study was mentioned in the previous section. For the Tier 1–2 method comparison, our statisticians set an ideal sample size of approximately 75 inventories from the full set of 750 to ensure sufficient representation in each of the 32 stratified sampling categories and to have approximately 12 inventories at each STAGES level.

For the stratified sampling, we randomly drew 12 inventories from each of the eight stages, or we drew all available inventories from the stage if fewer than 12 inventories were available. Within each stage, we attempted to balance the number of inventories across the four original CG/L scorers by randomly choosing from each original scorer’s collection of inventories. This would ideally result in three inventories for each scorer at a given stage. However, when this was not possible (due to insufficient number of inventories for some CG/L scorers), we selected all inventories from any scorer who had less than three inventories at the given stage and drew the remaining inventories (out of 12) from the remaining scorers. For these remaining scorers, the sample from each scorer was at least three inventories, and the specific number depended on the number of available inventories per scorer. For instance, there were no inventories for one of the CG/L scorers for Loehvinger's stage 3.5, but there were at least four inventories for each of the three remaining scorers; therefore, four inventories were sampled from each of the three remaining scorers. This stratification ensured a sufficient representation in the sample (n = 73) across the different stages and CG scorers.

**Selection of the STAGES scorers.** Four scorers were selected to perform the STAGES scoring for the 142 inventories used in this research. One was co-author O’Fallon, the developer of the STAGES framework, the only “master scorer” in this study, and the other three were the first certified scorers of the new STAGES system. These three scorers were of varied backgrounds, including a certified counselor, a lawyer, and a business consultant/coach trained in IT. O’Fallon has a background in elementary and special education teaching, school administration, and college teaching. She was the only scorer in this study who had prior experience scoring other developmental inventories—she had scored with both the Cook-Greuter system and the new STAGES system. All four scorers were familiar with developmental models prior to learning how to score. Therefore, there were four CG/L scorers and four STAGES scorers for this study. All 142 inventories were scored by three of the four STAGES scorers, and each of the 73 inventories in the Tier 1-2 study was scored by a varying one of the four CG/L scorers.

The STAGES scorers were randomly assigned to inventories such that two of the three less experienced scorers independently scored any given inventory (about 94 each) along with the more experienced scorer (TO) who scored every inventory (142). Each scorer worked independently scoring each inventory, blinded to the stage assigned by the other scorers. Each scorer was scheduled to score their own batch at about two inventories per week.

**3.4. Replication analysis of validity for Tier 1–2 data**

To establish validity of the STAGES developmental scores, we compared the single CG/L score of each inventory with the scores from the three scorers who scored that inventory using the STAGES model. The level of agreement for Tier 1–2 data was quantified by the weighted Cohen’s Kappa (κ) 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 also calculated the mean Kappa values across scorers.

**Kappa Statistic.** The weighted version of the Kappa statistic is commonly used to assess agreement for ordinal variables (such as stages of development). In weighted methods, a greater penalty is assigned to paired ratings whose scores are further apart. Throughout this article, we use the square method of weighting for all analyses when two or more scorers rate each inventory (Cohen, 1968). The square method is one of the common options for weighting mentioned in the Kappa literature. Square weighting also yields a Kappa value equal to the intra-class correlation coefficient under quite general conditions (Fleiss and Cohen, 1973). For paired ratings of each inventory, we used a weighted Cohen’s Kappa statistic, and for multiple raters scoring each inventory, we used the weighted Light’s Kappa statistic (Conger, 1980). The Light’s Kappa values can be directly interpreted as Cohen’s Kappa values (Landis and Koch, 1977). All calculations were carried out in R, version 3.0.0 (R Core Team, 2016).

---

10 It might be reasonable to assume that of the 20% who did not specify an education level, they achieved levels of education lower than the average.

11 Inventories were categorized into a table showing the CG-method scorer (four of them) and the CG score (eight levels through Strategist) for a total of 32 strata. Ideally there would then be three inventories by each of the four scorers for each of the eight levels. Ten of the 32 strata (defined by four scorers and eight stages) had zero inventories available, leaving 22 strata for analysis.

12 Here, STAGES levels 2.5 correspond to CG/L Diplomat, and 2.0 corresponds with the D3 “archived” inventories. Note that this is not a perfect match because the Diplomat inventories will contain both 2.0 and 2.5 because D3 was merged with Diplomat in the STAGES method.

13 O’Fallon, who scored all the STAGES inventories in this study, was also the scorer for 12 of the 142 CG/L inventories. It is unlikely that this overlap had any effect on the study outcomes because these CG/L scorings were done years before the STAGES system was developed, and O’Fallon stopped scoring with the CG/L method once the STAGES method was developed. As is seen later, analysis of the STAGES scoring was done for with and without O’Fallon’s scores.

14 The statistical weight by the square method is 1−{(i−j)/(k−1)}², where k is the number of categories and i and j are each the rank order of the categories chosen by two different raters for the same material.

15 A “<5.0” category is needed because although the sample inventories were expected to be at Tier 5, some inventories might have been scored by some scorers at a lower stage.
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, perfect agreement (Landis and Koch, 1977). We refer to the last category ($\kappa = 0.81$–1.00) as “very strong” instead of the commonly used “perfect” agreement since Kappa = 1.0 is the only value that indicates perfect (exact) agreement between the two sets of scores. Previous statistical studies in the Loevinger tradition tended to use correlation statistics (e.g. Pearson’s) to compare ratings, but, with the relatively limited number of stages used in these datasets, the Kappa statistics is more informative. Additionally, the Kappa agreement statistic appropriately penalizes a systematic bias of one set of scores versus another, whereas the Pearson correlation statistic remains unaffected by a constant systematic bias.

**Method Details.** Below we describe three details concerning the analysis.

1. **Scorer #1.** O’Fallon was in a unique situation: she was the creator of the model and the most experienced scorer. Moreover, she was the only STAGES scorer who had also studied the CG/L method and who had scored some of the original data using the CG/L method. Therefore, special precautions were taken to calculate results both with and without her STAGES scoring included in both studies (Tier 1–2 and Tier 3).

2. **Tier 1–2.** The STAGES model splits the CG/L Delta level into two levels: 2.0 and 2.5, complicating the CG/L comparison. To account for this, we ran comparisons in two ways: with stages 2.0 and 2.5 combined and with levels 2.0 and 2.5 separated. STAGE 2.0 in the Cook-Greuter scoring was initially recognized by Loevinger (called D3 or “Delta/3”). This stage was eventually subsumed in CG/L under the Diplomat stage because there were not enough distinguishing differences between the two stages and because there seemed to be less data in the D3 stage. The STAGES model not only requires a re-separation of these stages (based on its repeating structure) but also clarifies the distinguishing characteristics of each (i.e. passive vs. active mode).

3. **Tier 3.** At Tier 3 levels, the correspondence between the CG/L and STAGES levels is not one-to-one. Therefore, an inter-rater analysis for Tier 3 scores was conducted using STAGES scoring only. We performed the Kappa analysis using two different samples of inventories, Set A and Set B. For Set A, we compared inventories rated above 4.5 by any STAGES scorer, and for Set B, we compared inventories rated above 4.5 by all STAGES scorers. Moreover, in running the Kappa comparisons for Tier 3 we combined STAGES scores below 5.0 into one category, “<5.0”, resulting in five ordinal categories (<5.0, 5.0, 5.5, 6.0, 6.5).

### 4. Results

#### 4.1. Comparison of Tier 1–2 scores between the CG/L and STAGES systems

The estimated weighted Kappa values for the agreement of the CG/L score vs. STAGES score for Tier 1–2 (levels 1.0 to 4.5) are shown in Table 2. The results can be summarized as follows:

- When stages 2.0 and 2.5 were combined yielding seven distinct stages, we found very strong agreement between CG/L and STAGES scores for each of the four scorers ($\kappa = 0.81$–0.94).

### 4.2. Inter-rater reliability study of Tier 3 data and for all data

IRR of the STAGES scores for the Tier 3 ratings (5.0, 5.5, 6.0, 6.5, including a “<5.0” category) is shown in Table 3. Stages below 5.0 are combined into a single category, yielding five categories. As mentioned previously, two methods were used where Set A includes inventories for which any STAGES scorer assigned a stage in Tier 3 (n = 71), and Set B includes inventories for which all STAGES scorers assigned a stage in Tier 3 (n = 51) (therefore, Set B is a subset of Set A). The weighted Light’s Kappa statistic was used for multiple raters on each inventory. The overall agreement among the raters was “substantial” (ranging from 0.63 to 0.68) for three of the four analyses reported in the Table. The agreement was “moderate,” $\kappa = 0.56$, for the analysis of Set A with the more experienced Scorer 1 excluded.

**IRR of the full model.** The IRR among scorers across all 12 stages was **very strong** whether a) all scorers were included (Kappa = 0.82) or b) when the more experienced Scorer 1 was excluded (Kappa = 0.81).

To summarize, Table 3. Reproducibility of Tier 3 (stages <5 combined into one category).

| Scored for Set A (n = 71) | Scored for Set B (n = 51) |
|--------------------------|--------------------------|
| Scorer 1 excluded        | 0.56                     | 0.63                     |
| All scorers              | 0.65                     | 0.68                     |

16 All of the statistics are comparing the scorer to the expert. We cannot compare scorers to each other, i.e. obtain a multi-scorer inter-rating, because each scorer’s list of final 10 surveys is unique, i.e. they did not score the same surveys.
method, we can summarize other indications of its validity and reliability, which are described in more detail in Murray and O’Fallon (2020 to appear). First, we should note that arguments for the validity of STAGES rest substantially upon the strong results of the over 400 studies of the WUSCT mentioned above. Given that Cook-Greuter’s system is essentially the same as Loevinger’s, with the addition of a level at the top, and that our study shows substantial concordance with Cook-Greuter’s method, we can argue that the strong prior findings on the internal validity, face validity, construct validity, and internal validity of the sentence completion test continue to apply (though this ascription is more speculative for the top levels added on after Loevinger, which constitute a small percentage of the population). The face validity of the SCT continues to be demonstrated through the modifications made in STAGES, at least anecdotally, as subjects who use their assessment scores in conjunction with coaching or consulting services consistently report that the measurement both fits and deepens their self-understanding. Also, STAGES has been successfully used as a developmental assessment in about a dozen studies in various application areas, investigating things including organizational change in successful organizations, developmental analysis of women leaders, reflective self-knowledge in health care practitioners, psychological resilience in prison inmates, and assessing the sophistication of climate change understanding (see Murray and O’Fallon, 2020 to appear).

**Internal consistency.** Using a different data set than the one used in the primary study, a set that consists of all assessments scored using the STAGES model over approximately 10 years, we can use both classical test theory and item-response theory to evaluate the internal consistency of the 21 test items. Across the entire item pool of scores from 36 items, the Cronbach’s alpha statistic is 0.97 (i.e. ‘excellent’, from George and Mallery, 2003). Analysis of the instrument at the item level using both IRT and Rasch analysis also indicates that the assessment is very robust (Murray, 2020, to appear). This is consistent with prior research on the sentence completion test at the survey (ogive) level, for versions used by Loevinger, Cook-Greuter, and Torbert (Murray, 2019).

An additional indication of the strength of the sentence completion test method comes from assessing the internal consistency of newly created stems (sentence starters). As mentioned above, Loevinger and Cook-Greuter (and Torbert as well) maintained essentially the same set of stems for their studies (once each system was finalized); in part because modifying and validating the scoring procedure to include new stems was quite labor-intensive. Based on this one could argue that the validity of the method applied only to the specific stems used. STAGES, being theory-driven, does not have these limitations, and O’Fallon has developed about half a dozen alternative (or ‘specialty’) inventories in which 36 thematic stems have replaced 6 of the original stems. Internal consistency of these new items always show high internal consistency and high correlation with prior items. For specialty inventories on the themes of leadership, education, climate change, and love, the internal consistency of the new items by themselves is in the ‘good’ range (i.e. 0.8 to 0.9), while the entire inventory including the six new stems maintains a very high (.95 or higher) internal consistency (see O’Fallon and Murray, 2020 to appear).

**Longitudinal analysis.** Using a more recent database, we have also analyzed subjects who have taken the assessment more than once, to derive longitudinal measurements of validity. Evidence that each subsequent test is highly likely to yield a score equivalent or higher than the prior score (i.e. monotonic growth) is considered very strong evidence for a construct being “developmental.” Of the 1245 surveys in the database there were 143 that were re-tests, representing 115 clients; 88 of whom had taken one retest, 20 taking 2 retests, 5 taking 3 retests, and 3 taking 4 or 5 retests (the few re-test that were less than 3 months apart were excluded). The average time difference between re-tests was 2.1 years. In this analysis we ignore the time differences between tests (in future analysis we will also factor in retest gap time using multilevel modeling).

If we treat each of the 143 re-tests as an independent event: 38% stayed the same, 50% increased, and 11% decreased. Thus 89% increased or stayed the same. The 11% that decreased is acceptably explained by a combination of factors and “noise” including: rater error, test-retest variability (i.e. that tests taken even on the same day have some percentage chance of differing), or actual “regressions” due to serious life challenges resulting in cognitive or emotional stressors. Gains could potentially be attributed to test “practice effects,” but the 2 year average separation makes that very unlikely. That is, these results constitute substantial evidence corroborating prior research showing that the ego development construct is developmental in nature—now shown for the STAGES model.

Many of the subjects entered a program aimed at personal/professional growth (called “Generating Transformational Change” GTC) that included developmental models as part of the curriculum. It is possible that they learned vocabulary that lead to an increase in their (verbal/ textual) SCT score without advancing their deeper “enactive” development. (There does not yet exist research, or assessment tools, or even an adequate theory, allowing one to separate the verbal-only vs. non-verbal components of developmental change). If we focus only on the 47 retests from non-GTC subjects, we still see that only 17% of the retests lead to a decrease in scores, still substantially confirming the developmental nature of the ego development construct (the GTC cohort did improve more overall, with only 8% of retests decreasing).

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. Here we can add strong evidence for the developmental sequencing of O’Fallon’s newly defined highest stages. Of 84 retests in 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. This is an even stronger finding of monotonic sequencing than that for all three tiers together. The evidence is quite strong that for each metaware stage, it arrives longitudinally in the order expected (i.e. after the prior stage and before the succeeding stage).

**Additional Inter-rater scoring.** We can look to a different data source to confirm the high inter-rater reliability given in our main study reported in this paper. STAGES scorers are “certified” after completing a training program and practicing their skills until they achieve greater than or equal to 85% correct scoring (as compared with a master scorer) for 10 consecutive inventories in a row.\(^{17}\) This is for agreement at the inventory level. To obtain an additional indication of the inter-rater reliability of the scoring method we can assess the *item-level* (stem level) agreement. We have data for the 5 most recently certified scorers (trained over the last three years). Among these scorers, for their final 10 pre-certification scores, the survey-level accuracy (for the aggregate score over 36-times) was extremely high (much higher than the 85% minimum requirement, which may be changed to 90% based on these results). Of the 50 surveys (10 each for five scorers) only one did not have perfect accuracy (the one incorrect result was one level off)—thus the overall accuracy at the survey level was 98%.\(^{18}\)

At the item level, agreement was also excellent. The average accuracy was 93%. When taking the average accuracy over their 10 scores, the highest was 97% and the lowest was 88%. Looking at the set of 50 scores, the lowest accuracy was 72%, and the highest was 100% (the vast majority of errors were off by one level). Four of the five scorers had at least two of the ten surveys at 100% stem-level accuracy. This is very strong

\(^{17}\) We use percent accuracy rather than Cohen’s Kappa here. When comparing raters, the Kappa statistic may not convey the intended information when accuracy is very high (see Feinstein and Cicchetti, 1990, RE the “Kappa paradox”).

\(^{18}\) New sentence starters do need to be verified using statistical comparison with existing completions, and this can be done with about 20 inventories compared to the hundreds of inventories required to gather enough data to compile a representative set of examples for the CG/L manual for a new stem (for example, see Minaert, 2009).
evidence for the reliability of the scoring method (which, comparing these numbers to the main study in this paper, has improved in recent years, probably due to improvements in the training program).

5. Discussion

The STAGES framework adds both a new underlying structured model (with repeating patterns over 12 stages) and a new scoring method (based on general heuristic principles instead of example-matching) to prior frameworks for ego (or meaning-making) developmental assessment. In this article, we have described the model, discussed how it was created, and reported studies demonstrating its validity. STAGES successfully replicates the prior Cook-Greuter/Loevinger scoring system to a “very strong” degree up to level 4.5 (Strategist), and overall, it shows superior IRR as validated with four independent scorers. It contains modified definitions of the highest levels (the MetAware Tier), and we expect these definitions to continue to evolve as we learn more. Furthermore, the model’s structural patterns partially mirror respected contemporary developmental frameworks in the Neo-Piagetian tradition (Commons, 2008; Fischer, 2008), and STAGES serves as a theoretical link between this tradition and the construct-developmental tradition of Loevinger, Kegan and colleagues.

5.1. Benefits of the STAGES scoring system

There are several benefits of using the STAGES scoring approach over previous models and methods using the Sentence Completion Test to measure development. First, the STAGES scoring system uses a general set of scoring heuristics that apply to all stages and all sentence completions. Therefore, the new scoring system substantially reduces the effort it takes to change sentence starters when modifying the set of sentence completions in an inventory. With the STAGES scoring system, one can score immediately with a new sentence starter by using the three questions (based on theoretical principles) noted in our methodology. The theoretical principles, once learned, eliminate the need for a manual of sentence completions—a manual that is very time-consuming to create. Additionally, theoretically, it can be used to score text outside the sentence completion paradigm, including arbitrary essay questions or even books. This application is being experimented with but has not been formally evaluated. Moreover, it should only be used to score the developmental level of a particular text production and not to infer the developmental level of a person.

Second, embedding the structural logic of the person perspectives and repeating parameters supports clear definitions of each level and deeper understanding of the basic mechanisms driving development both for scoring and for individuals’ understanding of their developmental journey. The meaning imbued in the perspective definitions (parameters) support more enduring categories than exemplar categories and they highlight repeating patterns. For instance, 1.0 in the concrete tier provides a mirror for 3.0 in the Subtle Tier and 5.0 in the MetAware Tier. At each of these levels, a new self-identity arises. This kind of upshift arises for each of the four levels in the Concrete Tier, the Subtle Tier, and again in the MetAware Tier. Self-identity, cognition, focusing (attention), and awareness trajectories are measured with these repeating patterns. Furthermore, if an individual is in a receptive modality in one of the tiers (see Table 1), we know that the individual is likely to move to an active modality, e.g., moving from stage 3.0–3.5. Consider another example; individuals operating from an Individual perspective (third or fifth person) will still have their orientation to others grounded in the earlier Concrete perspective (second or fourth person).

Combining the first two benefits, 1. using a general set of heuristics and 2. the heuristic attributes as person perspectives each of which have distinct but interlocking definitions,—provides a novel and useful way to assess the development of cognition and consciousness.

5.2. Limitations of this study

There are many limitations and caveats involved in using developmental models and measurements (e.g. see Stein and Heikkinen, 2009; Murray, 2011). Reducing human meaning-making to a single ordinal scale (“center of gravity”), though useful in many ways, involves a rather blunt abstraction and simplification. Additionally, such written assessments might favor those who are more verbally articulate (as one might infer from the examples in Appendix 2). Moreover, the definitions and implications of the measured construct (e.g., ego development) are somewhat imprecise—users are advised not to pigeonhole or draw definitive conclusions about the characteristics of individuals or groups measured at a particular stage (Cook-Greuter, 2013; Murray, 2017). The additional limitations of this particular study have been presented below:

1. The proposed scoring system is a novel one, and this is its first scientific study. More studies are needed to support and extend this research. Typically, it takes more than one study to convincingly validate a new developmental framework.

2. This paper describes the only research on the three later stages (5.5, 6.0, and 6.5) to date (Cook-Greuter’s empirical research covered up to Construct Aware, 5.0) with all scores later than that held by a “Unitive” container; further research will be needed to continue to document and verify these later stages and to strengthen replicability as new data is input.

3. The participants for inventories in our data pool (including participants in the Generating Transformative Change (GTC) program at PI, consulting, and scoring contracted inventories for other entities) might have distinctive characteristics that render them non-representative of other more extensive populations of adults; we will need to compare this data set with inventories gathered from other population pools to claim better representativeness.

4. This study validates the STAGES scoring method. The scoring method is intended to reflect the STAGES model and its structural parameters. Therefore, we tentatively claim that the validity of the scoring system supports validity of the model. However, both the scoring manual and scoring skill are complex, and it is possible (though not expected) that scoring involves some decisions that are not direct reflections of the STAGES structural model. A more detailed analysis might thus be required to adequately assert that a study validating the empirical scoring method also fully validates the theoretical aspects of the model.

There is much to explore, and the field is still in early stages in terms of mapping out how people develop into the latest (“post-autonomous,” “second tier,” or “MetAware”) stages of development in human meaning-making. A theory mapping out a scientifically valid sequence of capacities beyond, say, the Strategist (4.5) level does not imply that it is capable of capturing all the aspects of development beyond that stage. Different theories might emphasize different emergent capacities. There are some indications that O’Fallon’s description of development after Tier 2 diverges, slightly but consequentially, from Cook-Greuter’s (in part due to distributing data across four stages instead of two) and that the two models disagree on certain characteristics of very late stage development. Without more empirical research of the CG/L model on these later stages, no absolute inference can be drawn. We believe that the STAGES model captures the same territory as the CG/L model with more clarity and explanatory power. Presently, this is a theoretical argument and the empirical question of whether it measures the same phenomena or something slightly different is an open question requiring probably a replication study for Tier 3 (similar to what was done for Tier 1–2).

10 Dissertations that have been done through universities include Prescott University, Saybrook University, and Williams James College. Pending dissertations are being done at Fielding University, and the University of Oslo, Norway.
**5.3. Future research**

Ongoing and future research using the STAGES model includes the following:

1) Continued verification of the STAGES scoring approach by cross-comparing the data from this study with data from different participant populations;

2) Continued verification of the IRR and developmental progression of the latest levels of development (MetaAware Tier) through both additional data collection and new types of populations;

3) We are developing “specialty inventories” that include a subset of stems focusing on specific life themes (such as leadership, psychology, parenting) by changing up to six sentence starters to reflect the new theme; we are in the process of checking the psychometric properties of these modified inventories, and will publish those results when complete.

4) We are engaged in research using artificial intelligence to automatically score the Sentence Completion Test so that this type of developmental assessment can be utilized on a sizeable scale for organization-wide and population studies too expensive to score by hand (as mentioned in Murray, 2017).

5) Several PhD and graduate students are using (or have used) the STAGES model in research on various aspects of human development, and these studies might inform the validity of the STAGES model and lead to new discoveries about development.

6) The fact that STAGES is based on a domain-independent model of language structure (as opposed to being exemplar-based) allows us to explore scoring text other than stem completions of Loevinger-style inventories. We have begun to experiment with scoring other types of text, for example, news articles, speeches, books, and social networking posts for developmental levels. In such works, it is the text (written performance) that is being scored and not an individual. This work is exciting but very preliminary

7) With the aid of our statisticians, we are re-evaluating the Ogive cutoff method that has traditionally been used to aggregate the scores of the 36 sentence stems to produce the center of gravity score. We are investigating whether more modern statistical methods including Rasch analysis (Rasch, 1980; Bond and Fox, 2001) will yield more reliable total scores (or sub-scores).

**5.4. Research involving human participants**

This study was reviewed by the Western Institutional Review Board and declared exempt under these citations: “We believe that the research fits the above exemption criteria. The aspect of the research where subjects will be completing the Sentence Completion Test, levels 10–12, is exempt under b(2) and the resoring of existing sentence completion tests, levels 1–9, is exempt under b(4).”

**Declarations**

**Author contribution statement**

T. O’Fallon: Conceived and designed the experiments; Performed the experiments; Contributed reagents, materials, analysis tools or data; Wrote the paper.

N. Polissar, M. B. Neradilek: Conceived and designed the experiments; Analyzed and interpreted the data; Contributed reagents, materials, analysis tools or data; Wrote the paper.

T. Murray: Contributed reagents, materials, analysis tools or data; Wrote the paper.

**Funding statement**

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

**Competing interest statement**

The authors declare no conflict of interest.

**Additional information**

Data associated with this study has been deposited at https://osf.io/k7pyf/.

---

**Appendix 1. Development of the STAGES Model and Scoring Rules**

Author O’Fallon trained as a scorer in the CG/L method and did scoring and related consulting for about six years. Based on her exposure to well over a thousand CG/L inventories, she began to hypothesize that the three dimensions of Tier, Individual/Collective, and Active/Passive could be used to explain the sequence of development described in the CG/L model. The STAGES model and assessment were then developed through a “grounded theory” approach (Haig, 1995; Hussein et al., 2014) that iterated between model refinement and data exploration until the model stabilized to its current form. O’Fallon then validated the model using empirical and psychometric methods described in this paper. The model is still considered new and is meant to be modified with new evidence over time.

O’Fallon worked with several collections of scoring data in creating the new model and scoring system (here we describe data used for model creation, not validation, which is described in the Methods section). The most valuable datasets were embedded in the three original Cook-Greuter and Loevinger scoring manuals (Hy and Loevinger, 1996; Loevinger and Wessler, 1970; Cook-Greuter, 2008). These manuals held over 25,000 examples of sentence completions that were already categorized (scored) by developmental stage. These were not whole inventories, but examples of sentence completions at each level of development, organized into about 10 thematic categories within each sentence stem and developmental level. Sentence completions in these manuals were scored and rescoring three times in the effort to sift the data through the lens of the new scoring method.

As O’Fallon was categorizing examples from these manuals, she was also drafting and refining general scoring rules as would be needed to instruct others how to score reliably and accurately. The next step was to score whole inventories to see if the scoring system would rate the final scores the same as scoring by the CG/L method. O’Fallon spent two years scoring inventories side-by-side with both methods (the CG/L scoring manual and the STAGES scoring method) and gradually refined the scoring rules of the STAGES scoring system to provide more accurate matching to the CG/L final score.

---

20 This Ogive method contains a number of assumptions and approximations which we will discuss in a future paper. What is important here is that the method is the de-facto standard that has been used in the numerous research studies using egodevelopment assessment for almost 40 years.
The original Loevinger manual provided practice for people to learn to score by themselves without a training class. The later scoring versions of the Loevinger test, including the Cook-Greuter approach (with higher stages incorporated), like the STAGES model requires taking an extensive scoring class to learn accurate scoring because scoring the later stages takes more training. Certified scoring usually requires the scorer to be aware of their own developmental level and to post inter-rater discussion emails to the scoring community on completions that they find difficult.

O’Fallon, due to the focus of her own research studies and the nature of her organization’s consulting work, had a higher number of very late stage inventories than did Cook-Greuter in her research data set. For refinement and triangulation of evidence supporting her definitions of late stage development, O’Fallon also studied late stage descriptions of scholar-adepts and sages including Sri Aurobindo (Aurobindo, 1992, 2000). Thus, though direct comparison of the two systems at these higher stages is difficult, O’Fallon believes that, as additional late stage data is obtained for both models, her system which divides these later level responses into four stages vs. two provides more refined distinctions of human development in this territory. As we discuss in the conclusion, for the MetAware Tier, it is not clear yet whether, from a theoretical stance, the two models have different ways of pointing to the same developmental territory or whether they diverge by referring to different territories.

Appendix 2. STAGES Cutoff Values

The Table below is used to determine the final derived score. The STAGES model uses the identical approach utilized by Loevinger (J. Loevinger, 1998) and updated by Cook-Greuter (1999) but, as described previously, adds “2.0” and two late stages.

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

Note how the levels being tested start with the highest (6.5), move down to 3.0, and then start at the lowest and work up to 2.5. This is because, according to Loevinger, the cutoff numbers were determined using a method inspired by Bayes Theorem, and their derivation takes into account how much evidence is needed to make a confident judgment (Hy and Loevinger, 1996). The 2.5 (Diplomat) level was estimated by Loevinger (and Cook-Greuter) to be the most prominent in the normal population and thus the most likely given no additional evidence. This is why 2.5 is the final or “default” value in the procedure.

Cook-Greuter used the cutoff of four for the two levels above Strategist (4.5), and STAGES duplicates that approach, using 4 for all stages above 4.5 (see Table 4).

The cut point criterion in the higher stages 5.0, 5.5, 6.0, 6.5 is analogous to the cut points in the Cook-Greuter scale but with more stringent scoring rules with each later stage.

Appendix 3. Scoring samples from the Concrete, Subtle and MetAware tiers

Below are several scored sentence completions as examples of the scoring process using the three questions. Note that within the scoring manual and procedure, there is more to the scoring than is depicted in these sentence completions, which are provided to give an overall orientation to the scoring process by providing a specific understanding of how the questions work.

**Sentence starter 1:** “At times I worry about_; completion: “the future of my country.”

**Based on question 1,** “Is the response Concrete, Subtle or MetAware?” We determine that this is a completion in the Subtle Tier because the word “future” is a subtle notion as no one can exactly visualize what will happen in the future (as opposed to a concrete idea such as, “I know that spring follows winter”).

**Based on question 2,** “Is the response Individual (is it all about me) or Collective (is it about we or us)?” We can see that this is about me and my worry, i.e. individual. This leaves two individual stages to choose from (3.0 and 3.5).

**Based on question 3,** “Is the response Receptive, Active, Reciprocal or Interpenetrative?” As an individual completion, our choices are either receptive (passive) or active language. The word “worry” is the verb. It is in the sentence starter so unless there is a verb written in the completion we need to use

### Table 4. Cutoff rules for a 36-item sentence completion test*.

| (Do steps in order and stop where true)                              | Assigned stage |
|---------------------------------------------------------------------|----------------|
| 4 or more rated at                                                  | 6.5 or higher  |
| 4 or more rated at                                                  | 6.0 or higher  |
| 4 or more rated at                                                  | 5.5 or higher  |
| 4 or more rated at                                                  | 5.0 or higher  |
| 6 or more rated at                                                  | 4.5 or higher  |
| 9 or more rated at                                                  | 4.0 or higher  |
| 14 or more rated at                                                 | 3.5 or higher  |
| 17 or more rated at                                                 | 3.0 or higher  |
| 7 or more rated at                                                  | 1.0 or lower   |
| 7 or more rated at                                                  | 1.5 or lower   |
| 7 or more rated at                                                  | 2.0 or lower   |
| - if none of the above, use                                         | 2.5            |

* Starting at the highest stage, 6.5, and proceeding down the table, assign the stage from the first row where the rule applies.
that verb. It is active so the best choice is the active quality and we have the final textual scoring of Subtle, Individual, Active, i.e. 3.5 late second person perspective or “Achiever”.  

**Sentence starter 2:** “Change is__; completion__ “inherent in living and connecting in the world; it prompts me to shake myself off from the slumber of consistency and embrace the excitement of newness.”

Based on question 1: “Is the response Concrete, Subtle or MetAware?” We determine that this is a completion in the Subtle Tier because “prompting” and “consistency” are subtle notions as no one can exactly visualize these terms.  

**Sentence starter 2:** “Is the response Individual (is it all about me) or Collective (is it about we or us)?” We can see that this is about we and connecting with the world. So, a collective score is required rather than an individual one. This leaves two collective stages to choose from (4.0 and 4.5).  

Based on question 3: “Is the response Receptive, Active, Reciprocal or Intepenetrative?” As a collective completion, our choices are either Reciprocal (foregrounding receptive/passive language) or Intepenetrative (foregrounding active language). The verb “it prompts” renders me a recipient of the action so it is passive; so, the best choice is the reciprocal/passive quality and we have the final textual scoring of Subtle, Collective, Reciprocal or 4.0, early fourth person perspective or “Pluralist.”  

**Sentence starter 3:** “When I get mad__; completion: knowing that uncontrolled unleashing of the power I now access can create undesired damage, I recognize the feeling tone in my awareness and I take myself on, stepping toward what brought on the anger, as I know that the emotion points at the growing developmental edge I have asked the universe to stretch.”

Based on question 1: “Is the response Concrete, Subtle or MetAware?” We determine that this is a completion in the MetAware Tier because s/he is recognizing their awareness (aware of aware). This narrows the choices down to the four stages in the MetAware Tier.  

**Sentence starter 2:** “Is the response Individual (is it all about me) or Collective (is it about we or us)?” We can see that this is about me and my anger. This eliminates all the stages that have a collective orientation, leaving two individual stages to choose from (5.0 and 5.5).  

Based on question 3: “Is the response Receptive, Active, Reciprocal or Intepenetrative?” As an individual completion, our choices are either receptive (passive) or active language. The verbs “knowing, access, create, recognize, take, stepping, know, points at, I have” are all active verbs. The best choice is the active quality so we have the final textual scoring of MetAware, Individual, Active, i.e. 5.5 Late fifth-person perspective or “Transpersonal.”  

**Sentence starter 3:** “Women are lucky because__; completion: “as the canvas they contribute to the Universes’ particular paintings in the Sacred’s art gallery by continuously receiving the brush of many colors, and thus are formed by the timeless, never ending layers of humanity’s pigment”.

Based on question 1: “Is the response Concrete, Subtle or MetAware?” We determine that this is a completion in the MetAware Tier because it foregrounds one whole/everything including timelessness. This narrows the choices down to the four stages in the MetAware Tier.  

**Sentence starter 2:** “Is the response Individual (is it all about me) or Collective (is it about we or us)?” We can see that this is about we and the universe. So, a collective score is required rather than an individual one. This leaves two individual stages to choose from (6.0 and 6.5).  

Based on question 3: “Is the response Receptive, Active, Reciprocal or Intepenetrative?” As an individual completion, our choices are either receptive (passive) or active language. The verbs “connecting with the world; it prompts me to shake myself off from the slumber of consistency and embrace the excitement of newness” are all active verbs. The best choice is the active quality so we have the final textual scoring of MetAware, Individual, Active, i.e. 6.0 early sixth-person perspective or “Universal.”

**Appendix 4. Per-Level Comparisons**

Table 5 shows a more detailed summary of alignment between the two scoring systems by level.  

This table was constructed by appending data sets with the scores given by all four STAGES scorers for the 73 inventories in the Tier 1-2 study. (As a reminder each inventory was scored by three of the four scorers, with Scorer 1 scoring all 73 inventories.) The total number of inventory scores is 218 across the four scorers, or 73 + 48 + 47 + 50. Also note that there are 7 CG/L column categories, and 8 STAGES row categories. This is because the data set used in the Tier 1-2 study was determined by the CG/L score being “5” or lower [i.e. 4.5 or lower in STAGES], while there were some cases [4] where an inventory scored as “5” in CG/L was scored as a “5.0” in STAGES scoring.

| STAGES | 1.0 | 1.5 | 2.0-2.5 | 3.0 | 3.5 | 4.0 | 4.5 | 5.0 | Total | Agrmnt | Agree +/- 1 |
|--------|-----|-----|---------|-----|-----|-----|-----|-----|------|--------|------------|
|        |     |     |         |     |     |     |     |     |      |        |             |
| 1.0    | 4   | 0   | 0       | 0   | 0   | 0   | 0   | 0   | 4    | 100.0% | 100.0%     |
| 1.5    | 5   | 5   | 3       | 2   | 0   | 0   | 0   | 0   | 15   | 33.3%  | 86.7%      |
| 2.0-2.5| 3   | 5   | **35**  | 9   | 1   | 0   | 0   | 0   | 53    | 66.0%  | 92.5%      |
| 3.0    | 0   | 2   | 9       | **21** | 1   | 0   | 0   | 0   | 33    | 63.6%  | 93.9%      |
| 3.5    | 0   | 0   | 1       | 4   | **28** | 7   | 3   | 0   | 43    | 65.1%  | 90.7%      |
| 4.0    | 0   | 0   | 0       | 0   | 3   | **19** | 6   | 0   | 28    | 67.9%  | 100.0%     |
| 4.5    | 0   | 0   | 2       | 0   | 3   | 9   | **23** | 0   | 37    | 62.2%  | 86.5%      |
| 5.0    | 0   | 0   | 0       | 0   | 0   | 1   | 4   | 0   | 5     |        |             |
| Total  | 12  | 12  | 50      | 36  | 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.)
As the CG/L system does not differentiate 2.0 vs. 2.5, these are combined. The original CG/L scores are listed across the top labeled by their STAGES equivalent with the CG/L ("MAP") level ID in parenthesis. The table 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."

Exact agreement percentages are shown between STAGES and CG/L codes for each row and for each column in Table 5. The table also shows Agreement "+/−1" (within one scoring level). The agreement percentages at the end of each row answer the question: what percent of the STAGES protocols scored at that level were correct, according to the CG/L scoring? The agreement percentages at the bottom of each column answer the question: what percent of the cases scored at that CG/L level were correctly scored at that level by the STAGES scorers?

From these detailed results we can conclude the following:

- Similar to the analysis of the Tier 1-2 study, the overall “weighted Kappa" score for this matrix is 76%, well into the “substantial agreement" level (given as (.61–.80) in Landis and Koch, 1977).
- The number of exact matches (values along the diagonal) varies per level. For levels above 1.5 there are substantial numbers of exact matches, but for the 1.0 and 1.5 levels, it would seem 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 are expected to score at these levels so practically this mismatch is not very important.
- The horizontal agreement in Table 5 varies from 52.8% to 70.0% over the levels above 1.5.
- The vertical agreement in Table 5 varies from 63.6% to 67.9% over the levels above 1.5.
- The “within one level" metrics are all quite high, from 86.5% to 100% for levels above 1.5.

(Note: There is no standard single metric that would summarize this overall matrix which sums (aggregates) over different raters. This is why our main analysis of the Tier 1-2 study shows a number of metrics including the Kappa values for each scorer that, together, describe the overall agreement (match) between the two scoring systems. In contrast, the point of the analysis in this Appendix is to compare the match between the two scoring systems broken down by each score level to check whether the overall strong average Kappa statistic given for the Tier 1-2 study might be hiding a poor match for a small number of specific levels.)

By summing over the four STAGES raters in the table above, we are, in effect, aggregating across them, establishing a marginal (overall) relationship. The table shows the aggregated matching between the STAGES rating and the CG/L rating.

Based on the data used in this study the STAGES model appears to bias the following levels a bit higher vs. CG/L: "2.0–2.5", 3.5, 4.0, while it appears to bias the following levels a bit lower than CG/L: 3.0, 4.5. There have been anecdotal concerns from colleagues 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 shows that there might be a small amount of such an effect in assigning CG/L Pluralist (4.0) scores into Strategist (4.5) but that this trend is reversed for the highest level where the bias is for STAGES to rate CG/L Strategists 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 a good comparison.
Kegan, R., 1994. In over Our Heads: the Mental Demands of Modern Life. Harvard University Press, Cambridge, MA.

Kohlberg, L., 1973. The claim to moral adequacy of a highest stage of moral judgment. J. Philos. 70 (18), 630–646.

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

Loevinger, J., Wessler, R., 1970. Measuring Ego Development 1: Construction and Use of a Sentence Completion Test. Jossey Bass, San Francisco.

Manners, J., Durkin, K., 2001. A critical review of identity development theory and its measurement. J. Pers. Assess. 77 (3), 541–567.

McChrystal, G.S., Collins, T., Silverman, D., Fussell, C., 2015. Team of Teams: New Rules of Engagement for a Complex World. Penguin.

Miniard, A.C., 2009. Construction of a Scoring Manual for the Sentence Stem a Good Boss—For the Sentence Completion Test Integral (SCTi-MAP) (Doctoral Dissertation. Cleveland State University. Cited 20 March 2018. Available at. https://engagedscholarship.csuohio.edu/cgi/viewcontent.cgi?referer=https://www.google.com/&httpsredir=1&article=1490&context=etdarchive.

Murray, T., 2011. Integralist mental models of adult development: provisos from a users guide. Integr. Leader. Rev. 11 (2).

Murray, T., 2017. Sentence completion assessments for ego development, meaning-making, and wisdom maturity, including STAGES. Integr. Leader. Rev. August, 2017.

Murray, T., 2020. Investigating the validity of the ogive method, including the use of Rasch analysis, for sentence completion test assessment for the STAGES model. Integr. Rev. 16 (1).

Murray, T., O’Fallon, T., 2020. Investigating the validity of the ogive method, including the use of rasch analysis, for sentence completion test assessment for the STAGES model. Integr. Rev. 16 (1).

Novy, D.M., Francis, D.J., 1992. Psychometric properties of the Washington university sentence completion test. Educ. Psychol. Meas. 52 (4), 1029–1039.

NSTA, 2011. Quality Science Education and 21st Century Skills. National Science Teachers Association 21 Feb. 2011. Available at. http://science.nsta.org/instaexpere t/PositionStatementDraft_21stCenturySkills.pdf.

O’Fallon, T., 2013. The Senses: Demystifying Awakening. In: Paper Presented at the Integral Theory Conference, 2013, San Francisco CA.

O’Fallon, T., 2011. STAGES: Growing up Is Waking Up—Interpenetrating Quadrants, States and Structures. In: Paper Presented at the Integral Theory Conference, 2011, San Francisco CA.

O’Fallon, T., Murray, T., 2020. Consistency studies for alternative sentence stem protocols for the STAGES inventory. Integr. Rev. 16 (1).

O’Fallon, T., Murray, T., 2013. The Senses: Demystifying Awakening. In: Paper Presented at the Integral Theory Conference, 2013, San Francisco CA.

O’Fallon, T., Murray, T., 2020. Consistency studies for alternative sentence stem protocols for the STAGES inventory. Integr. Rev. 16 (1).

Piaget, J., 1969. The Psychology of the Child. Basic Books, New York.

R Core Team, 2016. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. Available at. https://www.R-project.org/.

Rasch, G., 1980. Probabilistic Model for Some Intelligence and Attainment Tests. University of Chicago Press, Chicago, IL.

Rosenberg, S.W., 2007. Rethinking democratic deliberation: the limits and potential of citizen participation. Polity 39 (3), 335–360.

Scardamalia, M., Bransford, J., Kozma, B., Quellmalz, E., 2012. New assessments and environments for knowledge building. In: Assessment and Teaching of 21st century Skills. Springer, Dordrecht, pp. 231–300.

Selman, Robert L., 1971. The relation of role taking to the development of moral judgment in children. Child Dev. 42 (1), 79–91.

Stein, Z., Heikkinen, K., 2009. Models, metrics, and measurement in developmental psychology. Integr. Rev. 5 (1), 4–24.

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. Integr. Rev. 5 (2), 133–151.

Westenberg, P.M., Hauser, S.T., Cohn, L.D., 2004. Sentence completion measurement of psychosocial maturity. Compr. Handb. Psychol. Assess. 2, 595–616.

Wigglesworth, C., 2012. SQ21: the Twenty-One Skills of Spiritual Intelligence. BookBaby.com publishers.

Wilber, K., 2000. Integral Psychology: Consciousness, Spirit, Psychology, Therapy. Shambhala, Boston.