A Response to "An Examination of the STAGES Scoring System"

Tom Murray[1]

Appreciations

Below is a full response to Kristina Merckoll's "An Examination of the STAGES Scoring System," which is a critique of the STAGES validation study published in the Heliyon journal. First I want to thank Merckoll for the integrity with which he has engaged with us and the quality of his article.[2] We have been in communication about this research over several years, on and off, with many an exchange over emails. Though there have been differences of opinion, and we think for both sides there have been times when it seemed the other party did not well understand one's perspective, through these minor tensions Merckoll has engaged with great integrity and good faith throughout.

Also I want to acknowledge the high quality of what he has published, in terms of its attention to details and the rigor of covering several perspectives on the issue. Kristian has acknowledged that the statistical details and rigor of the original research publication were also exemplary. So our differences are more conceptual, theoretical, and perspectival. We can also acknowledge the "labor of love" effort he has put into this work. This type of academic work is rarely compensated for in the "liminal" fields of integral scholarly work, and the vast majority of our work on this research has also been "labor of love."

Overview of Our Main Counter Arguments

Though we will address specific points within the Critique paper below, the controversy can be captured simply. The following quote is Murray's attempt to summarize the main points of Merckoll's argument from an email [March 2, 2023], to which Merckoll responded "Your 4 points are correct." [April 19, 2023]:

> This is my summary of your thinking:
> 1. 85% exact agreement is seen as the industry standard for inter-rating reliability for SCT [Sentence Completion Test]

---

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

[2] This paper usually uses the pronoun "we" to include principles/authors of the original Heliyon study, specifically Murray, O'Fallon, and Polisar. Sometimes I use "I" to refer specifically to the author of this Response article.

2 This criterion should also be used to compare the two methods (what we called validation by replication)
3. Comparing STAGES with MAP using exact matching is much less than 85%, even for all 4 scorers, and even less if Terri is removed from the pool.
4. Therefore the two methods should not be said to correspond or replicate, statistically; and the "validity" argument for STAGES in the Heliyon paper is not supported.

Our strong difference of opinion is with the second premise, which forms the basis for conclusions #3 and #4. Based on our understanding of this research domain, and the knowledge of our two statistical consultants working on this research, we chose, and continue to argue for, the Cohen's Kappa statistic (in its "weighted" form) as the most adequate statistical measure of validity of this "replication" study. Put simply, if you use the Kappa statistic the results are great, if you use the exact match statistic suggested by Merckoll, the results do not look good! Now our task is to explain why the Kappa statistic is the appropriate choice in a way that the non-technical reader can appreciate. Other than this major difference of perspective, we will later discuss some of the secondary points made by Merckoll, and acknowledge that some of his observations are quite relevant and useful.

To clear one thing up at the start, we never claimed that STAGES scoring will give one a facsimile of a MAP score, or that it is a new or better method for getting your MAP score. Though in hindsight this fact could and should have been emphasized more in the Heliyon paper. Our research question is whether the two systems seem to be measuring close to the same underlying psychological construct, by comparing the final "center of gravity" scores (aggregated over 36 completions) from each system. Though the assessments are similar (using the sentence completion test), the STAGES model and thus the scoring methodology, are a radical innovation, and its correlation with previous methods for assessing ego development was not a given, but had to be demonstrated empirically.

The gist of our counterargument is as follows. The standard for 85% inter-rater reliability shared by most in the Loevinger tradition is a measure of "internal reliability" between raters *within* a particular measurement method. The main purpose of our study was to compare two different measurement systems and ask to what extent they seem to be measuring the same construct. This is an "external validity" task, which has a completely different analysis goal. Merckoll does not give a valid argument for why these two tasks must use the same criteria. We show that the kappa statistic is a much better criteria for an external validity task, and the exact match criteria is not even appropriate for a multi-scorer (4 scorer) study. The bulk of Merckoll's sub-arguments rely on this one assumption (#2 above), which we argue is false. So we stand by our original conclusions that the agreement between the two measuring systems is "very strong," ($\kappa = 0.81–1.00$) using common metrics from psychometric research (Landis and Koch, 1977).

Before diving into the details of our response we will address another important issue. Merckoll has shown that the inter-rating between the four scorers in our study (O'Fallon plus three others) is notably less than the common standard of 85% exact match used for professional services in the field. There are very good reasons for this, which we summarize here and expand later. These were the first three scorers that O'Fallon trained in her new scoring method, which was derived over time from the implications of her theoretical model. There was no completed scoring manual at

the time; and no scorer certification program. The training these three scorers received was the first iteration of any form of scorer training for STAGES, and a somewhat ad-hoc one at that. In fact, the very experience of performing this study helped inform what would eventually become the complete scoring method and manual.

O'Fallon faced a chicken/egg or bootstrapping dilemma here. There was little use in developing and running (and charging for) a scorer certification training if the scoring procedure was not validated empirically, yet to run the validation study inter-raters needed to be trained. So, the training was extremely minimal compared to what later emerged. Still, the inter-rater reliability based on the weighted Kappa statistic was quite good. That we got such strong Kappa validation metrics for the Heliyon study even with such "green" scorers is actually quite a *positive indication, since, were the study to be repeated with more fully trained scorers, the results should be even stronger*. In addition, we will show that the exact match criteria is inappropriate for a study that needs inter-rating between four scorers – which is an unusually high bar to set. (The kappa statistic can produce a 4-way match statistic among four scorers, which is not possible using the standard 85% match criterion.)

Upon getting preliminary positive results from her statisticians, O'Fallon commenced to develop a scorer certification program, and has certified over 20 scorers thus far (in two learning cohorts). We believe that our certification process, and the continued checking implemented for quality assurance, are at least as rigorous as that used by Cook-Greuter and associates.[3] Further, in our paper [Murray & O'Fallon, 2020] we analyzed the inter-rater reliability of the most recent cohort of scorers and found it to be well above the quality standard referenced by Merckoll (an average of 98% accurate at the inventory level, and 93% accurate at the stem level). Note that also in our use of Rasch Analysis for our full scoring database, the reliability statistic was also quite good, as reported in [Murray, 2020]. These are studies accessing our entire database of assessments, as opposed to the limited and early assessments that were the basis of the Heliyon study.

On "Validation" and "Replication"

A certain amount of this controversy is around whether our study "validates" the STAGES scoring system, "replicates" the MAP scoring system, or "proves" that STAGES is a new step in the ego development lineage. These are black-and-white terms that admit to significant ambiguity. In empirical research the question is usually "how good" or "how close" a match (correlation, correspondence, etc.) is, not a matter of either-or, true/false. Our statisticians recommended calling this a "replication study," i.e., one that tests whether a new method well enough matches an existing one to say they are measuring the same construct.

There are different interpretations of words like "replication" and "validation" within the literature, and we do not need to get bogged down in terminological nuances – we will let the statistics in our original publication speak for themselves. In many domains there are generally agreed-upon heuristics for a "good" match, and these differ across domains. As Merckoll notes,

[3] It is more rigorous than the MAP system was when O'Fallon was working within that system, though MAP procedures may have changed since then.

one needs to consider not only the empty statistics, but "what the numbers represent." For example, if one is doing a study for an intervention or technology where an error would lead to an unsuitable drug treatment or a failure of an automotive breaking system, the standards are higher than, say, in the fields of sociology and psychology where one is comparing groups or trends to make conclusions that are not life-threatening.

Study Purpose and Context: A Radical Proposal

Merckoll notes that it is "important to recognize ... that STAGES does not present itself only as a new developmental construct. Its stated objective is to update and strengthen a current one, both in regard to its underlying developmental theory as well as its assessment." This is mostly correct, but to be more precise, STAGES is meant to extend and improve the theory and measurement of the ego development construct, not the MAP scoring method itself. As noted above, we have never said that "if you receive a score on a STAGES assessment it should match well with what you would get on a MAP assessment." Though we intend STAGES to correspond well to MAP overall or on average in the subtle tier, we would never guarantee for any individual that the two scoring systems will correspond.

In fact, we know this to not be the case. As noted explicitly in the Heliyon paper (and elsewhere), the two scoring systems diverge at both the lower stages (Concrete Tier) and higher stages (Metaware tier), because STAGES has a moderately different understanding of what adult development means for those stages.[4] In the middle stages, which is where 80-95% of the population falls (depending on how one defines/estimates it) the measurements correlate quite well. But they are still not exactly the same, as the scoring methods and rules are *very* different (STAGES scoring is based on 3 "parameters" while MAP scoring is primarily in reference to a standard set of example sentence completions – as described in more detail in the Heliyon paper). The STAGES system is not another "rater" for the MAP system, it is a completely different technique for assessing what we hypothesized to be approximately the same psychological construct.

Construct indeterminacy. Merckoll says that "According to the integral principle of enactment, STAGES will never be an exact replica of ego development, since its methodology and theoretical framework is different." We agree completely that the two systems use different methods and thus are not measuring *exactly* the same thing. To illustrate the indeterminacy of a measured construct, there might be many ways to measure the "economic vibrancy" of a city, the

[4] For example, Cook-Greuter maintains that after construct aware (also called ego aware, or 5.0 in STAGES) responses become shorter, simpler—perhaps more poetic or mystical. O'Fallon's model claims that 5.0 is entry into a new person perspective (5th) and is thus receptive in nature—a process of learning about a new world of phenomena. STAGES predicts that following each receptive phase is an active phase, which is not necessarily simpler or more concise. Two longitudinal study reported in Murray & O'Fallon (2020) show this to be the case. As an example of where the systems diverge for the lower levels, the MAP scoring system scores responses illustrating profanity or outbursts of anger at very low levels ("impulsive"). We consider this a conflation of "shadow crashes" with developmental sophistication. We discuss any such shadow indication in the report given to participants, but score the completions based only on the three parameters. A 1.5 or 2.0 individual in STAGES is either at a very young age or has cognitive impairment, and shadow manifestation is treated differently.

"intelligence" of one's personality, or the "health" of an ecosystem. Statisticians (in our case psychometricians) invent standards for determining if two measurements can be argued to be measuring more-or-less the same construct.

Ego development is a psychological construct, like "intelligence," or "introversion," that does not exist as such "in nature" but is a human-invented concept pointing at something we intuitively sense exists. For example, there is no exact thing in the human mind/brain that is "intelligence," yet the concept points to an intuition we have about something so important about human capacities that many believe is worth measuring and even standardizing. There is no singe brief, operational, or standard definition for "intelligence" – the concept admits to a cloud of ideas having a conceptual "family resemblance" (Rosch & Mervis, 1975; Lakoff & Johnson, 1999). At bottom, scoring high on a specific "intelligence test" means one does well at the specific tasks on the test (with its specific items for logical inference, abstract thinking, etc.), and does so under test-taking contexts and pressures. Intelligence as a concept (a psychological construct and vague concept about human capacities) is much broader than performance on a particular assessment instrument. Two alternative test of IQ may predictably differ, but they may correspond sufficiently to conclude that they are both measuring essentially the same construct.

Likewise with ego development. No one has claim on the definition of ego, and not Loevinger, nor any other scholar, can or has given a definitive and precise definition of ego or ego development. But these concepts are important enough to our understanding of the human condition that we try to define and measure them.[5] (This is how research and theory work in all of the non-exact or non-physical sciences, and why the ideas of internal validity vs. external validity were fleshed out in the first place.)

Merckoll would seem to understand the above on the indeterminate nature of psychological constructs, as he says "there is no 'ego development' lying around waiting to be discovered, it is enacted and brought forth by the particular methodology." Yet we disagree around whether the STAGES model or assessment corresponds to the MAP system, given the differences. Cook-Greuter's and Torbert's systems were actually a significant departure from Loevinger's original work, and (from generally accepted community lore and anecdotal evidence) Loevinger was skeptical, if not downright critical and unconvinced, of this new direction, and did not see it as following directly in her footsteps at all, in terms of either theory or measurement. We seem to be at a similar juncture here with questions of whether STAGES is a valid successor, or legitimate branching path, within this lineage. And it is natural that when new methods are proposed there will be differing opinions about whether one builds validly upon, or improves, the prior.[6] Our stance is that the STAGES model does build directly upon Loevinger's and Cook-Greuter's and Torbert's work (and Wilber's theory), as do other projects around the globe.

Because O'Fallon was trained as a MAP scorer, we always knew that certain sentence completions will score differently between the two systems. And thus the total aggregate "center

[5] Personally (in Murray, 2023 to appear), in recent theoretical papers, I have argued that ego development is very close to what we might call "wisdom skill," and also that Kegan's theory of consciousness development is also very close to our understanding of ego development.

[6] Perhaps there is a parallel to the fact that Freud and his followers were not "on board" with Jung's work either.

of gravity" score over all 36 completions will also sometimes differ between the two scoring methods. It is a very complicated set of conditions that are encoded into each of the scoring systems (which is why one needs an extensive training to become a certified scorer in either system), and the systems differ in many ways. It was known to be possible that a person scoring 4.0 Pluralist in MAP might score at STAGES 3.5 (Achiever) or 4.5 (Strategist) in some situations, in edge cases where the differences in the theories was most pronounced.

Radical Proposals

The STAGES scoring system offers a *radical proposal*: that most of the complexity described in ego development methods is the result of just three underlying variables or parameters. Looking at Loevinger's original work, the often-cited description of the stages by Cook-Greuter (2005), or any number of other publications describing each ego development level, one sees that it requires paragraphs to describe or define each of the stages. There was, prior to STAGES, no concise way to describe each or any of them, never mind the trajectory through the entire sequence. (Though many authors provide terse one-phrase summaries of each level in figures and tables illustrating the levels.) The STAGES model proposes that each stage is adequately captured by just three parameters: Tier (C/S/M), individual-vs-collective, and passive-vs-active.[7] Though measurement in the "soft" human sciences is much less precise than in the hard sciences, this is analogous to how the periodic table of elements managed to describe what was previously a set of elements each having a complex set of descriptors (density, magnetism, reactivity, malleability, etc.) through a radically simple two-parameter (two-dimensional) table.

On the face of it, it would seem unlikely that something like this is possible in an area as complex as ego development (I myself was skeptical at the start). This had to be proven and the primary way to do this was to compare with another ego development scoring system. The STAGES model and scoring system were so radically different that even a "moderate" or "substantial" match would have been considered promising news – yet the match was "very good.".

In his later suggestions on improving the STAGES scoring system Merckoll says the "STAGES scoring system should not in any way be built upon the MAP manual of sentence completions. A STAGES scoring system should be built directly upon STAGES own stage definitions, without any implicit reliance upon the MAP manual." In fact, this is exactly what O'Fallon has already done, and the scoring manuals are radically different. For most of the stages the MAP manual gives long lists of example completions organized into themes, whereas STAGES is not at all based on exemplars, or even particular vocabulary or ideations, but rather on general rules and heuristics about sentence structure and deep form, stemming from a sophisticated understanding for the parameters.

O'Fallon arrived at her theoretical model after years of study in developmental theory, linguistics, and integral theories/philosophies (including Pearce and Aurobindo's work), including extensive dialogs with Wilber, and cross-checking her intuitions against a database of scored responses. After all of this work and revision, she was confident that she had something – yet it

[7] The active/passive and individual/collective parameters are sometimes combined into a "learning style" or complexity scale of receptive/active/reciprocal/interpenetrative (see the Heliyon article).

remained to be proven to work empirically, and to show that the system could be taught to others. The exact determination of the three parameters is non-trivial[8] and O'Fallon created a scoring manual (38 pages including figures and appendices) for training scorers.

As noted, using the Kappa statistic commonly used in such situations, the match was "very strong." As noted by Merckoll, if one uses an exact matching criteria that gives close misses no credit, the alignment is much worse. So which method is right?

Kappa: Exact vs. Weighted Matching Criterion; and Internal vs. External Validity

Merckoll's main argument is that because the generally accepted criterion for inter-rater reliability (a measure of internal validity) in the ego development lineage is 85% exact inter-rater agreement, which he terms the "basic inherent quality standard" for the domain, and that this criterion should be used in the study in question. We disagree definitively here. Inter-rater reliability is a measurement of how closely raters ("scorers" in our case) agree, or how closely any two measurements using *the same method* agree. It gives an indication of reliability and "internal validity" for a measurement, which indicates an assurance of similarity and repeatability over different measurements by different people or instruments at different times and over different individuals being measured.

Measures of *internal* validity say that an assessment is a solid reliable "measuring stick" but do not tell us whether the measurement measures what it is supposed to. Determining whether a measurement measures what it is presumed to measure is called "*external* validity" and specifically for us "*construct* validity;" and when construct validity is assessed by comparing two methods that are hypothesized to measure similar constructs, the term "*convergent* validity" is used.[9] In our case the question at hand, the one the Heliyon study was meant to answer, is whether the STAGES scoring method adequately matches the MAP scoring system used by Cook-Greuter and associates. The question is basically "is the STAGES scoring system measuring more-or-less the same psychological construct as the MAP?" *This is a completely different question than inter-rater reliability, so the validity criterion need not be the same.*

How does one test the hypothesis of construct "convergence" between two measurement systems? Our statisticians (one who is a senior practitioner in the field) determined that the *weighted Kappa* statistic was the best method.[10] The Kappa statistic is used in many types of comparison studies, is often used for inter-rater studies, and is sometimes used for convergent validity studies. It is considered a more rigorous method than alternatives because it takes into consideration that some scores might match simply from random chance. We can provide evidence

[8] For example, the "active" in the passive/active parameter includes active orientation to an object such as ownership/responsibility, as well as the traditional active/passive syntactic markers. Determining whether an object is Concrete, Subtle, or Metaware is also complex. In the individual/collective parameter, describing what "collective" means for abstract (subtle or Metaware) entities is non-trivial.

[9] See Wikipedia article https://en.wikipedia.org/wiki/Validity_(statistics).

[10] The "weighted kappa...assigns less weight to agreement as categories are further apart" (Viera & Garrett (2005) and is used in situations like ours where the categories are ordinal or scale-like.

from the literature that the kappa statistic is appropriate for convergent validity[11], while Merckoll gives no references supporting his opinion that it should not be so used, and provides no support from the literature that the quality standard for internal validity *must* (or even should) be used for external validity studies.

Merckoll argues for the 85% exact agreement criterion. *We argue that any exact match criterion (including a Kappa exact match) is inappropriate, and that a weighted measurement is most appropriate.* Exact methods are usually used for categorical values (e.g., red/green/blue/yellow; animal/mineral/vegetable), while weighted methods take into account *how* close (or far) the match is. For example, in an exact match system a scoring difference of 2 (e.g., between 2.0 and 4.0) is seen as the same error as the difference between of .5 (between a 3.5 and 4.0) – both cases are simply "wrong" or "non-matching." In a weighted system a difference of 2 is much worse than a difference of .5, and the amount of difference is factored into the overall statistic. Developmental *scales* measure gradations in a changing phenomenon, and the distinct "levels" are used for convenience and simplicity.[12] Merckoll's note that an interrater error of greater than one is problematic attests to the fact that this is a hierarchical scale, not a horizontal set of distinct categories.

It makes sense to use the more stringent criteria for inter-rater reliability in cases where individuals will be given their results, as is done in the businesses that use both MAP and STAGES scoring, because there are enough differences between successive stages that a higher degree of scoring confidence is warranted. But with more general research questions, such as our external validity study, the more stringent exact match criteria is not warranted.[13]

Merckoll gives "three fundamental problems with the use of Kappa values as constituting a validation proof:"

[11] Tests for convergent validity (comparing two methods or models, as opposed to two raters or measurement tools) sometimes use Kappa statistics but more often use correlations such as the Spearman because most such studies are looking at continuous variables, rather than ordinal or scaled categories. Still some internet searching easily identified three studies that explicitly use the kappa statistic to assess convergent validity: Cools et al. (2010); Burns et al. (2014), and Kahn et al. (2013). For example, in Cools et al.: "A Kappa correlation coefficient (0.67) provided moderately strong support for convergent validity" (p. 597).

[12] The question of whether the development of various skills happens smoothly/continuously or whether there are "jumps" or "gaps" between the stages is a controversial one. For simple tasks there is evidence that the subskills of horizontal development enter a chaotic relationship and then rather quickly coalesce into a higher order skill. For example, a child can practice riding a bicycle for weeks or months, and then, seemingly suddenly, one day they are riding the bicycle. But common sense observation tells us that for complex skills like ego development (and other "lines" such as relational, spiritual, and leadership development) people don't suddenly jump to an entirely new level in a short period, though they probably do have relative plateaus and rapid changes in the chaotic process of development.

[13] Merckoll says "a fundamental limitation of using the Kappa here, is that it does not properly account for 2 stage deviations." This is difficult to interpret because giving extra penalty to more serious errors is exactly what the weighted form of Kappa does. Perhaps Merckoll simply means that for trained scorers an error of 2 stages is completely unacceptable, and that the kappa does not account for this. But again we would argue that this "quality standard" is for rater certification to insure a quality product when assessments are sold, but is not germane to a convergent validity research study.

1. First, one cannot use those interpretations quoted above (“substantial”, “very strong” etc.) to directly validate a hypothesis.

2. Second, the Kappa is not a good construct to use when we take into account how the two scoring systems relate to each other.

3. Third, in this instance it can be shown that the Kappa statistic actually conceals the a priori quality standards the scoring systems are based upon.

His arguments for each of these are specious in our opinion. On #1 he states that "the linguistic descriptions are more like guidelines," which is certainly true. McHugh (2012) says "Cohen suggested the Kappa result be interpreted as follows: values ≤ 0 as indicating no agreement and 0.01–0.20 as none to slight, 0.21–0.40 as fair, 0.41– 0.60 as moderate, 0.61–0.80 as substantial, and 0.81–1.00 as almost perfect agreement" but goes on to note that the interpretation of the values depends on the domain in question. In fact we did not want to use the term "almost perfect" and chose to use the term "very strong" agreement. McHugh suggests that studies that yield recommendations for medical interventions should err on the side of caution, and for the "clinical laboratory ... many texts recommend 80% agreement as the minimum acceptable interrater agreement."

So, yes these indicators such as "substantial agreement" are relative to the domain. But we are not doing high-stakes medical research here, and our use of Cohen's language aligns with many other studies in psychometric, psychological, and educational assessment. As explained in more detail in the next section, the reader can look at "the whole picture" and decide for themselves if the match is "close enough." But that is subjective for each reader and it is appropriate to refer to a quasi-standard interpretation of "good" match, as we did. Merckoll's argument is that the "very strong" agreement cannot be correct because his preferred method of exact match does not yield a very strong agreement, but this logic works backwards from a desired conclusion, and we have argued that there is no defensible argument why the more stringent criteria is necessary.

As to point #2: Merckoll claims that "the STAGES scoring system is a wholly derivative method, and a high correlation, or correspondence, is already a given." This is a spurious argument. First, as noted, STAGES is a radical departure in terms of model and methodology. Second, this line of reasoning would seem to invalidate any study that attempts to improve upon or extend a prior measurement, for example a new method for scoring introversion/extroversion. Such studies begin with a common foundation between the two methods and an intuitive confidence that they will correlate at least minimally, and the question is how tight the correlation is. To say that some degree of correlation is expected ("a given") does not detract from the "very strong" match we conclude in our construct validity study. In fact, the greatest use of the Kappa statistic is in inter-rater reliability studies, where, yes, one expects there to be a significant correlation between any two trained raters, but the question is about "how close" they are.

As to #3, we have argued at length about why the "quality standard" referred to by Merckoll applies to inter-rater reliability and scorer certification, but not to the convergent validity research we are doing. This quality standard was first established, and rightly so, by those, including Cook-Greuter and Torbert, who first took Loevinger's assessment out of its academic research setting

and charged people money for an assessment and debrief. In research studies, where individuals are anonymized and one is looking for general trends rather than precise individual measurements, the confidence interval of any given measurement need not be as stringent.

In sum, if one insists that exact match is the only way to go, then one must agree with Merckoll that STAGES does not agree very strongly with MAP. But we argue that there is no good reason to limit the comparison in this way. There is no defensible reason to constrain comparison to exact match, and there is no defensible reason to say that the method used for internal validity (reliability/inter-rating) *must* be used for external validity (construct/convergent validity). Without the need for these constraints, we have chosen what we argue is a perfectly appropriate statistical method. So as to Merckoll's claim that "the STAGES scoring system is not fully validated as an alternative to the MAP," we agree that it is not, nor was it ever intended to be, an alternative way to score a MAP inventory; but STAGES *is* an "alternative" model and scoring system for the construct of ego development – and our interpretation of the results, especially in the context of the other validity studies we have done (Murray & O'Fallon 2020) supports the idea that the model and scoring system are "valid" or "validated," to the extent that one single study can do so (and in Murray & O'Fallon 2020 we report on additional validation research).

In his section on "the role of the statistician" Merckoll argues that our statistician was hampered by the fact that he is not an expert in developmental assessment (the "lead statistician behind the study does not appear to approach the situation from an adequate understanding of all the relevant aspects of the scoring system. He does not consider the quality standards in the scoring systems, or how these two scoring systems already relate to one another"). Our statistician was well aware of the 85% match criteria for inter-raters and decided not to factor that into the analysis as it was not necessary (and in fact inappropriate for a multi-scorer study, and described below in the inter-rating section). In hindsight we could have avoided this aspect of the controversy if we had addressed it directly in the Heliyon paper. In fact our statistician *is* an expert at psychological (psychometric) analysis,[14] and there is *nothing special about developmental* psychometrics here (as opposed to, say, personality or instructional design assessment). The question at hand is quite general: when conducting an external validity study, is one constrained to use the quality standard accepted for internal reliability in that domain? We can find nothing in the literature that supports this premise. Calling the internal validity criteria a "quality standard" does not change this, as it is a quality standard for internal validity only.

Statistical Stories

The exact match percentage, the correlation, and the kappa statistic are statistical aggregation methods that summarize a complex "story" about a data set into a single value. We need such summarizing statistics (like mean, median, and mode) to make simple global comparisons such as "how well do two data sets correlate or match each other?" But much of the nuance of the story and the "territory" is thus lost. A picture is worth a thousand words, and a table is probably worth a hundred statistical summary values. That is why most research papers include graphs, plots, or tables as well as the final statistical aggregations. We copy Table 5 from the Heliyon article (and

[14] For example, Polisar has over 200 peer-reviewed articles in scientific journals, and has been asked to be an expert witness regarding statistical methods for numerous lawsuits and legal proceedings.

Merckoll's paper) below. It was included in the original article so that "the whole story" was transparently visible.

Table 5. By-stage agreement between STAGES and CG/L scoring.

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

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

Visual inspection of Table 5 shows that there is a pretty close match between the STAGES and MAP scoring systems. As noted by Merckoll "A Kappa based on these 146 data points is 0.84 ... within the 'very strong' agreement category. The reason for this relatively high Kappa number is that the data are clustered around the diagonal..." This is exactly right, as for a study asking whether STAGES is measuring more-or-less the same construct as MAP, we want to give some credit to "close but not exact" correspondences.

As noted elsewhere, we would expect to get an even better match, and much less variability between STAGEES scorers, if the study was repeated with the more experienced scorers we have today – but the match would never be perfect because the two scoring systems do differ on some details. Of course, if one counts only exact matches (the diagonals in red) things look worse than if one looks at adjacency and general trends. But looking at the whole picture the two systems clearly track quite well. A visual "eyeball" match of "quite well" is not a scientific or rigorous evaluation of such things, so one must turn to statistical norms to quantify the match. Using our weighted Kappa method the match is within the range of *very good* based on widely accepted ways of interpreting Kappa values.

In the Heliyon paper we go into greater detail on the implications of the mismatches observed in this table. For example, though STAGES seems to score some levels on-average higher than MAP, other levels are scored on-average lower – but there does not seem to be an overarching bias either way.

Looking at this table, there do appear to be a few troubling data points with larger mismatches (differences greater than one level). One could ask "what is going on there?" – why were some

scores so different between the two systems? Some of these differences are actually predicted, for example some "shadow crash" phenomena that MAP would score at 1.0 or 1.5 might have a level of complexity that STAGES would score much higher. As another example, Cook-Greuter's system expects 5.0 (Construct Aware) and above to usually be simpler or more elegant than 4.5 (Strategist), but the STAGES model predicts that there is a stage after 5.0 (i.e. 5.5) that will include more "active" and complex language. Thus for completions that have some Construct/Ego Aware (5.0) characteristic yet are also complex or lengthy, the MAP scorer has little choice but to score it at 4.5, while the STAGES scorer uses their discernment to decide whether such completions are 4.5 or 5.5.

Though the question of "*exactly, stage by stage,* what are the qualitative and quantitative difference between MAP and STAGES scoring?" is intriguing, answering this was not the purpose of the replication study, and would constitute a major study in itself (requiring significant funding or pro-bono time investment). We cannot draw these types of conclusions from the Heliyon study data, as there is no way to tell how much of the variation is due to explainable differences in the models/methods, vs. how much is due to scorer "error" or subjectivity – a topic to which we turn next.

Scorer Inter-rating – Then and Now

Here we move from discussing external/construct/convergent validity between two completely different measuring methods, to discussing inter-rater reliability (internal validity) *within* a scoring method. As noted it is very important to distinguish between these two goals within psychometric analysis. Though Merckoll seems to partially conflate them by insisting that the same standard be used for both. We have already shown why the weighted kappa statistic, rather than the exact match criteria, was chosen to compare the two rating systems. Our statistician also chose the kappa statistic for inter-rating the four STAGES scorers. We will discuss (1) why the kappa statistic was chosen for IRR, and (2) why the inter-rating appears low using the exact match criteria.

Merckoll is correct in pointing out the inter-rater reliability of the scorers in this study was not up to the 85% exact match criteria accepted in the field (and the additional qualification of all rater differences or errors being no greater than one level). As we have already noted:

1. The three scorers (other than Terri) were the first to be trained, and that training was ad-hoc without a scoring manual or certification program.

2. This is because of the bootstrapping issue: there was no need to develop a manual and certification program (both quite sizable tasks) until and unless the basic method was proven to be valid empirically – which was the purpose of the replication study.[15]

[15] Loevinger used a similar bootstrapping method in developing her, then novel, scoring system. There were many iterations of "blind" studies followed by vigorous detailed discussion about the difference of opinion on scores, with the method and scoring manual being iteratively improved along the way. Loevinger (1998, p. 5) describes the early process of defining the levels: "because we initially had no scoring manual, we discussed as a group how to classify each completion, *trying to imagine* the type of person who would give such a response" (emphasis added). This led to the first of a series of scoring manuals, all of them exemplar-

3. The scoring method and its documentation evolved concurrently with the Heliyon study, and with the input and experience gained as the 4 scorers worked together on this study.[16]

4. The 85% exact match criteria for IRR was instituted (after Loevinger's work) when the SCT moved out of a purely academic research context to be used commercially to provide individuals with a score (and report and debrief). There is always assumed to be some "error" or uncertainty in human scoring, and quantitative research studies averaging results over many (anonymous) participants can tolerate much more measurement uncertainty than the commercial context.[17]

5. The larger margin of error (inter-rater variability) in the Heliyon study actually bodes *well* for the replication study, as we achieved a "very strong" kappa result even under these conditions, and would presumably get even more favorable results with more highly trained scorers.

6. Currently all of our approximately 20 certified scorers have shown excellent inter-rater reliability (beyond the industry standard). (And the three scorers who participated in this study also went on to become fully certified.)

Given this context, the entire section on "Discussion on Why the Independent Scorers do not Score Accurately Enough" is moot and/or misinformed speculation. Statements such as "Over time, the effect of the lack of adequate scoring precision will become apparent in any population being scored" are true but irrelevant here. Merckoll also speculates upon scoring issues specific to the Metaware level, which we cover later.

Why use Kappa for IRR?

The argument for why the Kappa statistic was used for inter-rating the four scorers is different from why it was used to test convergent validity between the two scoring systems. O'Fallon wanted the study to be very rigorous and sought out a senior statistician for this job, and described her desire for rigor. The statistician suggested an inter-rating between *four* raters, which is considerably above and beyond the standard used in most inter-rating studies, and well beyond anything done previously in the Loevinger tradition. Each rater has some "error" or subjective variability in their scoring – either through actual error (human mistake), category wiggle-room in the scoring rules, edge cases in the data, or having a unique and valid subjective interpretation of

based, i.e., a completion is scored by matching it to a set of real examples (which, in later manuals, are grouped into categories based on thematic similarity).

[16] Proper method of course dictates that scorers produce their scores without prior discussion of each inventory. However, once the official research data is recorded, scorers can and do have inter-rater conversations about the items they scored differently. This method is used universally within the SCT assessment community—less experienced scorers thus learn 'on the fly' and continuously improve, and the expert scorers also learn from the experience, as new perspectives are taken into account, and complex cases and grey areas are excavated to improve the articulation of the methodology for future iterations of the scoring manual.

[17] The Central Limit Theorem in statistics states that random measurement errors increasingly cancel each other out for ever larger sample sizes.

the scoring rules that differs from others. Given this inevitable variation in scores it is *much* more difficult to get four scorers to agree than two (i.e., the probability of difference increases multiplicably with each additional scorer). So, as Merckoll has noted, the experimental design itself is quite rigorous, in fact unusually so.

Importantly, the exact match criteria is not appropriate for more than two scorers (as explained above). Surely one is not going to count only the items where *all four* scorers agreed. There are complex ways that the exact match criteria can be used to compare every possible 2-way combination of each of the scorers, and then combine this information, but this sort of thing is exactly what the Kappa statistic already does. Specifically, there is a variation of the Kappa statistic for multiple raters called the Light's Kappa statistic (see the Heliyon paper for references), which also has a weighted form.[18]

Given that exact match is not a good match for IRR in this study,[19] there may be other statistics that could have been used, but the kappa is standard and sufficient. It is not difficult to find support for this in the literature. For example, in their paper "Understanding interobserver agreement: the kappa statistic," Viera & Garrett (2005) explain: "diagnostic tests often rely on some degree of subjective interpretation by observers. Studies that measure the agreement between two or more observers should include a statistic that takes into account the fact that observers will sometimes agree or disagree simply by chance. The kappa statistic (or kappa coefficient) is *the most commonly used* statistic for this purpose" and "kappa can provide *more information than* a simple calculation of the *raw proportion* [i.e., percent] of agreement" (p. 360, emphasis added).[20] Kappa values are not directly comparable with percent agreement and are often lower for a given study. For example, McHugh gives an example where "the percent agreement is 0.94 while the Kappa is 0.85."

O'Fallon's Greater Correlation with MAP

We appreciate Merckoll's Figure 2 and 3 analysis showing O'Fallon's scoring alone and the other three raters in a table, which illustrates how much better O'Fallon's scoring corresponds to the MAP system. Though our original paper spoke at length on how the Kappa scores were very good even with O'Fallon removed, we did not take the analysis into this direction. The differences shown by Merckoll are considerable. But what can be concluded from this? Merckoll suggests that it shows that O'Fallon's scoring was overly influenced by her training in the MAP model (vs. the other scorers who did not train in the MAP system). Though some degree of this prior influence is

[18] Statisticians will also be familiar with Fleiss' kappa, an alternative used for multiple raters.

[19] The study was designed to treat all scorers as potentially equal, that is, it the study design assumed that they all had sufficient expertise. Of course, they were not equal, and O'Fallon was the top expert. But the study design was not interested in comparing each scorer against an expert, for this the exact match criterion would have been appropriate. Rather, in acknowledging that O'Fallon was in a special situation because of her prior training as a MAP scorer, we ran all statistics both with and without O'Fallon included, as discussed later.

[20] Similarly, from McHugh (2012): "The concept of 'agreement among raters' is fairly simple, and for many years interrater reliability *was measured as percent agreement* among the data collectors.... *Cohen's kappa... is a robust statistic* useful for either interrater or intrarater reliability testing... As with all correlation statistics, the kappa is *a standardized value* and thus is interpreted the same across multiple studies" (emphasis added).

possible and even likely, one cannot determine how much influence this had from the data. As we have noted, it is equally possible that the difference is because the 3 scorers received only minimal scoring training (and had never previously trained in *any* scoring system; while O'Fallon had significant experience in both systems).

We also appreciate Merckoll's Figure 3 showing the differences between O'Fallon and the other scorers, a rendering we did not include in our analysis. He illustrates how much differences (which he calls "random noise") exist between the three new scorers, and between all new scorers and the master scorer. Again, we expect these differences to be much smaller for our more recently trained scorers (and even for the original 3 scorers who have now have undergone full certification and have years more experience). In fact, it is notable that Merckoll *did not find any systematic pattern* (such as a consistent offset) in the differences between the new scorers and master scorer, but rather that the differences seemed "random." Random differences argue for scorer inexperience as being the main issue; whereas a systematic difference would argue more strongly that O'Fallon had brought some bias from her MAP scoring into her STAGES scoring.

In sum, it is reasonable to surmise that the significantly closer correlation to MAP scores that O'Fallon achieved vs. the other scorers was due to some influence (or unconscious "bleed-over") from her prior training as a MAP scorer. This possibility is exactly why the Heliyon study showed results with and without O'Fallon, and, unsurprisingly the match without O'Fallon was lower. Yet the match without O'Fallon was still "very good" according to the kappa statistic. Using the exact match as a criterion exaggerates and highlights O'Fallon's unique position, but does not change the results of the study.

On "Long Term Consequences" and Scorer Certification

In his sections "The STAGES Certification Process vs. the Replication Study," "Improving the STAGES Scoring System," and "The Long-term Consequence of too much Scoring Error," Merckoll speculates at length upon what the Heliyon study, completed when the STAGES model was in its infancy, implies for the larger context of the STAGES scoring method as it exists today, 10 years later. (The actual scoring and data collection was done in 2012-2014, though it took until 2020 for the analysis to be complete, the paper written and submitted, and finally published.[21])

We have shown why it is quite invalid to extrapolate from this early study, which used ad-hock scorer training, to later contexts following the establishment of a rigorous certification program. In addition, Merckoll was aware that we have published more recent studies of inter-raters (and other validity studies, see Murray & O'Fallon, 2020) and he explicitly declined (in email messages) to read and consider these additional studies. This exclusion is acceptable (though still regrettable) for the sections of Merckoll's critique that limit discussion to the Heliyon findings, but is less defensible as he ventures beyond to speculate about the STAGES scoring method in general as it exists today.

[21] Unfortunately, the process of submitting papers to peer reviewed journals and moving through the publication process can take much longer than expected. We applied to several journal before identifying the right one to use. Adding to this, in the middle of the period of 2013-2016 O'Fallon had to take considerable time off from the project for intensive caretaking of an ill family member.

Thus, almost all of what is said in these three sections is also based on assumptions we have argued are incorrect.[22] Below we will say more about our scorer certification and inter-rating process, so there is no doubt in there reader's mind as to its rigor. But first we think it is worth noting that, while we have exposed the data in this study to public scrutiny, and have also published other papers on the inter-rater reliability of STAGES scoring, we are not aware of any such data or publication centered on the MAP scoring system. Though both methods explicitly commit to the 85% exact match standard for scorer certification, we know of no published or publicly available study or data that verifies this for the MAP system and its trained scorers, which would allow a comparison of IRR of certified scorers between the two systems.

Also, Merckoll critiques the inter-rater correlations on a stage-by stage basis, first by differentiating out higher stages (Metaware) , and second my combining estimated population data with the results at each stage. No one has done this more refined by-level of analysis for MAP scoring IRR, and again, it is possible that if the same scrutiny were applied to the MAP system, the results would be no better than STAGES. In contrast, Torbert and colleagues *have* published some studies of IRR on their variation of the SCT (the LDP and GLP instruments), which are discussed in Murray (2017).[23]

One could argue that the performance profile of the MAP scoring system is not relevant to the discussion, but we believe that it is relevant because Merckoll brings in the comparison many times. Though the research base of the MAP system is not relevant to a discussion about the specific results, the comparison becomes relevant as one speculates more broadly, as Merckoll has done.

STAGES Scorer Certification

Because Merckoll extrapolates erroneously and at such length with speculations about the STAGES scoring system's quality as it exists beyond the Heliyon study, we need to set the record straight. Merckoll notes, "from a communication with Cook-Greuter," that "when someone during the [certification] exam misses by more than one stage, for the inventory as a whole, *'it is a serious concern'*." We agree that this level of error is a serious concern for certifying scorers for the commercial application of giving people (and charging them for) an individual score. STAGES scorer certification incorporates the same level of rigor (and possibly stronger, as noted above)

[22] We do recognize some potentially misleading text in our Heliyon paper, at "Till date, about 10 individuals have been certified to score using the STAGES model and more are under supervision for certification." Here we were mentioning that, while the scorers in the study were the first, and less rigorously trained at that time, *since* that study a rigorous system has been put in place (today the number of certified scorers has grown from 10 to 20). However, "till date" may not have been clear enough, and this text could have been interpreted to mean, incorrectly, that the three Heliyon scorers were likewise certified and fully trained prior to the study.

[23] Torbert & Livne-Tarandach (2009) report an IRR of 69% perfect matches and .90 within one level, and a Cronbach's alpha measure of internal consistency of .91. Torbert (2014, p. 7) reports an inter-rater study of "805 measures, each of which could have been scored at 13 different levels [using early- and late-specifications of the developmental levels]. The result showed a .96 Pearson correlation between the two scorers, with perfect agreement in 72% of the cases, with a 1/3 action-logic disagreement in 22% of the cases, and with only one case of a disagreement larger than one full action-logic."

To become certified as a STAGES scorer one must achieve *100%* accuracy on 10 test inventories in a row. Any error means the scorer must start over again to reach the 10 in a row. In addition is a requirement of 85% accuracy at the *stem* level. This means that, for the 36 completions over the 10 test inventories, 85% accuracy is required. 85% accuracy at the stem level is *much* more rigorous than 85% accuracy at the full inventory level. In Murray & O'Fallon (2020) we analyzed the five most recently certified scorers at the stem completion level, and found agreement to be above the target level of 85% exact match at the stem/item level: the average accuracy over the five was 93%, with the highest at 97% and the lowest at 88%.

Merckoll is correct to note that adequate scorer inter-rating accuracy upon completion of a training program does not necessarily imply that the scorer maintains that level of accuracy going forward. STAGES International incorporates continuous quality assurance tests in an ongoing way for its scoring work. This is done in two forms. First, a master scorer does regular inter-rater spot-checks on all scorers, and more frequently on those who are less experienced or who have not been scoring much recently. This ongoing supervision includes getting feedback from the expert scorer on errors discovered. Second, *all* stem completion scores in the Metaware tier must be inter-rated by another scorer. This is done either through the scoring platform, via a Slack discussion app, or via private email messages. The purpose of this inter-rating is not to calculate an IRR statistic, but to cross-check with another scorer to minimize the chance of bias or error for these more difficult and rare later stages (5.0 stem scores are actually not that rare amongst the data in our database for clients that choose to take the STAGES assessment). Third, in order to maintain one's status as a certified scorer, one must continue to score and be inter-rated, and participate in professional development and review seminars.

The STAGES scoring certification process takes about a year to complete with weekly study assignments. As noted, the scorers who participated in the Heliyon study had nothing even closely approximating this level of training (they had *less than one day* of introduction to the methodology, based on preliminary scoring rules); and the scoring manual was not yet created. Subsequent to the Heliyon study all three scorers went through the equivalent of the full certification program.

Considering Real-life Stage Distributions

Merckoll's paper explores at length what the implications are when one combines the study results with estimates of how developmental stages are distributed over the population. These distributions are speculative because they were taken from small studies and we actually have no large scale studies for concluding large scale distributions; and also these were done some decades ago and it is difficult to say how culture has evolved. Still, until we have large scale studies, most of us working in this field do refer to the small-sample-size studies that have been done as we try to tell a story about the implications of developmental theories for society.

But even if the population estimates *were* more valid, the conclusions are moot based on our argument here. They all start with the assumption that the exact match criteria is the only way to compare the two systems, and continue to show that the inadequate match is even worse if one takes these distributions into account. But we have argued that the starting premises are invalid.

To the Heliyon quote: "Overall, there does not appear to be a significant shift, higher or lower, in STAGES vs. CG/L scoring," Merckoll responds that "it does not take into account the distribution of the developmental stages in the actual population. In table 5 each of the levels 3.0, 3.5, 4.0 and 4.5 have 36 observations each. This equal number of scores for these different stages is a good idea when it comes to the actual study but cannot be used to check the practical effects of mismatches in any actual population." We do agree that it is very useful to use population statistics to gauge the impact of research results on actual populations. But that was not a goal of the original study. Merckoll's point that errors in more frequently seen levels have greater potential impact is reasonable, but the inter-rater match from the Heliyon results do not generalize to overall trends because of the minimal scorer training we have mentioned.

Metaware Dreams

Merckoll includes a section focusing on the Metaware tier. As before, he uses the exact match criteria created for the commercial use of providing individual scores as his benchmark – and we have already described why this is inappropriate. We agree that the IRR exact match found in the Heliyon study is not as rigorous as what would be needed to certify raters to score in the Metaware tier, and above we describe the quality of our actual certification process and also the extra steps taken to insure rigor of every Metaware stem scored for a client. Still, as noted in the paper, using the standard kappa statistic, the IRR for scoring at the Metaware tier was "moderate to substantial," and thus quite good given the indeterminacies of the new scorers and scoring system already mentioned.

Beyond these statistics, Merckoll speculated on the "existence" of the later stages in the STAGES model. First, we are all on the same page in knowing that the categories described by developmental models are human-created constructs that do not exist as such "in reality." Merckoll says of the Metaware stages: "As I see it, we are more in the realm of hypothesizing about the real nature of these later stages, unlike the earlier ones which are previously studied much more extensively." This is certainly true, as this is newly charted territory. Here we are in a situation more like when Loevinger first created her theory and scoring method, with nothing well-established to compare it to. In fact, no one else is doing serious research on these later levels, and as far as we can tell STAGES has the most robust data set for doing research in the later levels, not to mention a model that articulates more structure and nuance beyond Strategies/4.5, vs. alternative models.[24]

While noting the novelty of research into Metaware, Merckoll ignores that fact that we have two additional published studies (summarized in Murray & O'Fallon 2020) on the Metaware tier. In Murray & O'Fallon (2020) section "6. Longitudinal Analysis" gives evidence that, indeed, the iterating patterns described by the STAGES model can be seen beyond Strategist/4.5. And Section "7. Analysis of Late Stage Patterns" provides additional evidence of the validity of STAGES model into the Metaware tier.

[24] In Ken Wilber's The Religion of Tomorrow (2017), he cites O'Fallon's STAGES model (p. 236 and 517) as "one of the few researchers actually investigating 3rd-tier levels".

Merckoll says: "As for testing at the MetAware Tier, I am still in the area of curiously not knowing. I am inherently skeptical of the notion that we can use language as the sole means to score at these later stages, especially the very latest ones. In any event, the ego's cunning way of portraying itself into a transpersonal realm is perpetual and will never end." Again, we agree here. There is much to learn about the later stages of ego development, and the depths of one's consciousness and skill set probably become ever more difficult to assess though a written assessment the later the stage.

What we and others in developmental research do know is that language can become both more sophisticated and more diverse and unique as the stages progress. From this we can speculate that the example-based method does better the lower the developmental levels, and ever worse for later levels, because it becomes ever more difficult to capture the range of possible expressions (responses to the sentence stems) with a finite set of examples. Even when completions become more terse and poetic, they can still be unique and surprising.[25] In fact, the MAP scoring system successively phases out example-based scoring and moves to rule-based scoring for its latest levels.

Merckoll speculates that, because it is based on just three parameters, that people's prior understanding of the model may influence their text responses, resulting in inflated scores. As our scorers are trained to recognize superficial attempts to guess language that would match with the parameters, we believe this is not a major concern. As noted, though the nature of the three parameters can be described in a few sentences, it requires a 38 page scoring manual and a year's training to learn the complex considerations involved in scoring the responses. The question of whether prior knowledge of developmental models might unintentionally, or intentionally (as in "gaming the system"), affect scores an interesting empirical question for both STAGES and MAP, which both use rule-based methods for the latest levels. Studies in the Loevinger tradition indicate that the sentence completion test is resistant to such biases.[26] There has been concern that the effect is worse for later stages, and our study in Murray & O'Fallon (2020) section "7. Analysis of Late Stage Patterns," and the accompanying paper "Deconstructing Developmental Constructs" was designed to address thisissue,empirically (though only partially). The data showed evidence that one cannot score late in the STAGES model simply by using spiritual or esoteric language that fits the simple description or state experiences associated with late-stage development. Again, more research is warranted.

We did the research, and even though one study allows only limited conclusions about the larger picture, it is empirical work making reasonable conclusions. In trying to critique the STAGES

[25] Merckoll's statement "it is definitely much more difficult to internalize the scoring method for the MAP; that would require extensive study of the scoring rules" is puzzling given how little Merckoll knows about the STAGES scoring system and manual. Also, this statement is in the context of a section on scoring the Metaware tier – and it is exactly here that the MAP system diverges from the example-based method used by Loevinger and veers into more rule-based methods for scoring.

[26] Redmore's (1976) paper "Susceptibility to Faking of a Sentence Completion Test of Ego Development" says: "Effects of faking on a sentence completion test of ego development. explored in five experiments, partially confirmed the developmental hypothesis that persons can lower their ego development test scores but not increase them. When faking high on retests. subjects' scores either stayed the same or increased by about half a step. Only intensive study of ego development seemed to genuinely increase scores" (p. 607).

model and scoring system in Metaware, Merckoll uses phrases such as "we certainly cannot rule out that the difference in the scoring approaches can have something to do with..." This is weak argumentation compared with the bulk of the article, as it focuses attention on "what cannot be ruled out" even for something them could be very unlikely. These more speculative assertions would seem to be evidence that the author is stretching to make a pre-determined point rather than offering solid evidence.

Conclusions

We believe we have addressed and countered all of Merckoll's major critiques. Most importantly: (1) his insistence that the 85% exact criteria be used; (2) his concerns about sub-standard inter-rater reliability and what that portends for STAGES in general, and (3) his speculations about the indeterminacies of our model for the Metaware tier. To which we have responded: (1) explaining why the exact match is inappropriate (as it conflates internal vs external validity criterion) and why we chose the weighted kappa statistic; (2) the scorers for this first study, and only for this first study, were minimally trained, and still the results were very strong; and (3) we have additional published studies arguing for the validity of the STAGES model and scoring in Metaware.

Though earnest and respectful, our debate has not been without its pokes and intellectual "noogies," and for both parties the writing hints at moments of consternation with the other, as befits an energetic disagreement. Merckoll's analysis, which we argue is based on incorrect premises, casts a very critical, and we think undeserving, shadow on the STAGES assessment method overall (reaching beyond the Heliyon study). Merckoll was aware that several other studies of STAGES were completed following the Heliyon study but decided not to read them prior to writing the Critique. These include longitudinal studies, a Rasch analysis, and inter-rater reliability studies, each of which adds additional support to the model's validity (though this set of studies is not a comprehensive picture, and more can be done, *see* "Summary of STAGES Validation Research," Murray & O'Fallon, 2020). We wish that Merckoll had read about these studies before submitting his critique, which ignores the implications of the other studies to paint what we think is an unfairly negative picture.[27]

Merckoll suggests that there is some confirmation bias or other bias at work in the STAGES methodology: e.g., "why has this simple truth about the STAGES scoring system apparently gone under the radar by the authors of the article?" We believe we have been properly scientific in our methodology and conclusions. Given our roles and the history of this and the prior critique published, we do not believe that any of the parties is a fully objective participant in this debate. In trying to enact a more "integral" or wholistic debate process I will reflect on possible outcomes of my own biases, which have been revealed thanks to Merckoll's dedicated work on this project.

[27] As it took some time to get the Heliyon paper published, we were able to reference these later studies in the Heliyon paper. We believe that in combining these studies that there are more empirical studies supporting that validity of the STAGES model than exist for Cook-Greuter's MAP method. Though the MAP instrument has been used in many studies, we are not aware of any published studies that attempt to validate aspects of the MAP scoring system beyond Cook-Greuter's original dissertation work.

For example, I notice that I did not insist that we also compare our kappa values for inter-rating using the exact match method (for the inter-rating aspect of the paper, this is not about the convergent validity aspect). We did not do the extra work to notice that the IRR of our Heliyon scorers was *that* much worse than our own standard for scorer certification. Merckoll also illustrated the tighter match O'Fallon has with MAP vs the other scorers, another revealing analysis we could have run. Well done empirical research is costly and time-consuming, and, often raises as many questions as it answers – yet one has to wrap it up at some point. When we received the statistician's report that the results looked very strong, we were guilty of being relieved and celebratory, and anxious to move on to publication – yet we could certainly have done additional inspections of the data.

Merckoll says "If this present article can spark a fruitful and respectful dialogue where we approach the topics directly, I will have achieved my aim." I hope we have done justice to this wish. The care, detail, and integrity of Merckoll's analysis has caused us to look more deeply and ask ourselves hard questions and has given us an opportunity to clarify things that prior publications, or "word on the street," left open to misinterpretation. Even if we disagree with many of his main points, we appreciate the open collegial intention, not to mention the patience, he has brought to our conversations over several years. The field of adult development, with its close association to spirituality and "consciousness raising" themes, attracts many "true believers" and includes a good share of claims made without reference to empirical study, and also contains too little in the way of well-intentioned critical debate – so Merckoll's challenge, based largely on a statistical analysis, is health for our field in general.

The STAGES model is indeed relatively new, and a few studies, promising as they may be, do not close the book on such a complex area. Though it is difficult for "independent scholars" not employed through universities or large grants to find the time and resources to do in-depth empirical research, we do intend to continue our research in adult development. Advances in scaled-up scoring will allow us to research trends for large populations; there is still much to learn about the psychology and phenomenology of Metaware development; and there is the potential to do more comparative analysis between the many adult developmental models we are familiar with.

References

Burns, R. D., Hannon, J. C., Allen, B. M., & Brusseau, T. A. (2014). Convergent validity of the one-mile run and PACER VO2max prediction models in middle school students. *Sage Open*, *4*(1), 2158244014525420.

Cook-Greuter, S. (2005). Ego development: *Nine levels of increasing embrace*. Unpublished manuscript. Retrieved 1/20/22 from https://legacy.beingfirst.com/wp-content/uploads/2016/02/EgoDevelopment.pdf.

Cools, W., De Martelaer, K., Vandaele, B., Samaey, C., & Andries, C. (2010). Assessment of movement skill performance in preschool children: Convergent validity between MOT 4-6 and M-ABC. *Journal of sports science & medicine*, *9*(4), 597.

Jordan, T. & Murray, T. (2020). Deconstructing Developmental Constructs: A Conversation. *Integral Review, 16*(1). 568-592.

Kahan, D., Nicaise, V., & Reuben, K. (2013). Convergent validity of four accelerometer cutpoints with direct observation of preschool children's outdoor physical activity. *Research quarterly for exercise and sport*, *84*(1), 59-67.

Lakoff, G. & Johnson, M. (1999). *Philosophy in the flesh: The embodied mind and its challenge to Western thought*. New York, NY: Basic Books/Perseus Books Group.

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

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

McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia medica*, *22*(3), 276-282.

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

Murray, T. (2020). Investigating the Validity of the Ogive Aggregation Method, Including the use of Rasch Analysis for the Sentence Completion Test and the STAGES Model. *Integral Review, 16*(1) 395-440.

Murray, T. (2023, to appear). Wisdom, Complexity, and Transcendence: A developmental frame. Chapter in J. Stevens-Long & E. Kallio (Eds.), *The Oxford Handbook of Adult Development and Wisdom*. Oxford University Press.

Murray, T. & O'Fallon, T. (2020). Summary of STAGES Validation Research. *Integral Review, 16*(1) 39-68.

O'Fallon, T., Polissar, N., Neradilek, M. B., & Murray, T. (2020). The validation of a new scoring method for assessing ego development based on three dimensions of language. *Heliyon*, *6*(3), 1-15. https://doi.org/https://doi.org/10.1016/j.heliyon.2020.e03472

O'Fallon, T. & Murray, T. (2020). The STAGES Specialty Inventories: Robustness to Variations in Sentence Stems, *Integral Review, 16*(1). 441-456.

Redmore, C. Susceptibility to faking of a sentence completion test of ego development. *Journal of Personality Assessment 40*(6) (1976): 607-616.

Rosch E. and Mervis, C. (1975) Family resemblances: studies in the internal structure of categories, *Cognitive Psychology 7*, 573-605.

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

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

Viera, A. J., & Garrett, J. M. (2005). Understanding interobserver agreement: the kappa statistic. *Fam med*, *37*(5), 360-363.

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

Wilber, K. (2017). *The religion of tomorrow: a vision for the future of the great traditions-more inclusive, more comprehensive, more complete*. Shambhala Publications.