An Examination of the STAGES Scoring System

Kristian Merckoll[1]

Abstract: This article examines two empirical studies which are referred to as validating the STAGES scoring system. The verification claims do not hold up under scrutiny, and the wrong conclusions were reached because the statistical calculations were not grounded in an adequate understanding of the relevant aspects of the scoring theory itself. This is shown by applying simple descriptive statistics that does not require any statistical background to understand. In the subsequent discussion on possible reasons for why the STAGES scoring system has too much scoring error, the conclusion is that a major root problem is that STAGES has defined itself as an extension of an existing developmental construct (EDT, Ego Developmental Theory). This notion must therefore be reconsidered. Standard calculations are provided to demonstrate effects of too much scoring error. The article further discusses the limitation in using only language as the means to assess the latest stages in the STAGES model. A distinction between the STAGES theory and the STAGES scoring system is maintained throughout, and the theory is not discussed or questioned directly. The position is taken that EDT and STAGES should continue to develop alongside each other as two related, yet different, developmental models.

Keywords: Anecdotal evidence, constructivist developmental models, developmental altitude, empirical evidence, integral enactment, quality standards, scientific rigor, STAGES, statistics.

Introduction

In the article *The validation of a new scoring method for assessing ego development based on three dimensions of language*, (O'Fallon et al., 2020) the authors present and describe a replication study. They draw the broad and generalized conclusion that *"This study validates the STAGES scoring method.*" They also present an inter-rater reliability study on the latest stages in the STAGES model. It may be helpful to read that article, but it is not necessary in order to follow my arguments below. You find a version of it here.

In this examination I will take a closer look at the way O'Fallon and colleagues have applied statistics in order to arrive at their validation conclusion. I will show that by digging a bit deeper into this material, it is imperative to reconsider the validation claims in the article. Contrary to validation of the STAGES scoring system, I will show that STAGES' own studies reveal the opposite: The STAGES scoring system does not assess ego developmental stages with sufficient

[1] **Kristian Merckoll** holds a master's degree in mathematical statistics from the University of Oslo. He has a background as a research scientist at the Norwegian Computing Centre, and now runs his own independent consulting business. He has studied advanced Integral Theory at Fielding University, is an Integral Master Coach from ICC (Integral Coaching Canada) and an Associate Lectical Coach (from Lectica). He is also an ordained priest in the Rinzai Zen lineage.
kristian@merckoll.no

scoring precision. I will make comments on what I view as implications of the analysis, and I invite anyone interested to contact me to discuss this or ask clarifying questions.

Background

A critique of STAGES, calling for empirical rigor, was published in 2017 in Integral Leadership Review (see Appendix). It was written by Susanne Cook-Greuter, Beena Sharma and Ken Wilber, and I was one of several co-signers. A rebuttal from STAGES proponents was published in the same ILR issue, and this rebuttal had many references to the apparent validation of the replication study which I am examining in this article (see Appendix). At the time of the rebuttal, the results from the replication study were not yet published, so those claims on replication could not be examined.

Terri O'Fallon later showed me a version of the STAGES validation article prior to its first publication (February 2020), and I had some comments on the use of AQAL as a framework to explain STAGES. I readily admit that I did not immediately notice what I later came to discover. I assumed the analysis was well done, as was also reiterated several times in the aforementioned rebuttal.

Later on, I went back to the STAGES validation article and did a deeper investigation into the STAGES scoring validity claims. What eventually spurred me to go into this in more depth were the specifics of the validation claim, and the way STAGES positions itself in relation to ego developmental theory.[2]

My initial review of the STAGES claim - including calculations and comments - was presented to the authors of the validation article (Terri O'Fallon, Nayak Polissar and Tom Murray) in an earlier version of this current article, in October 2020. I have now chosen to present my arguments to a larger audience, with additional calculations and comments. If this present article can spark a fruitful and respectful dialogue where we approach the topics directly, I will have achieved my aim.

My interest and motivation for this investigation is rooted in two decades of engagement with Integral Theory. I have been well exposed to the theories and scoring systems discussed in this article (ego development, MAP and STAGES). I am also familiar with the Lectical approach, and I have 30 years of practical experience as a professional statistician.

The audience for my article is whoever is interested in developmental psychology and its assessments, especially ego development and STAGES. This article is not written for academic publication, and the tone is therefore slightly more informal at times. Most importantly, I have put emphasis on using simple statistical language, and making all the calculations possible to follow

[2] A comment on notation: When I speak of ego developmental theory, or shorthand EDT, I am referring to the CG/L (Cook-Greuter/Loevinger) ego developmental theory. For the scoring system using the manual, as well as the result called the MAP (Maturity Assessment Profile) I will use the term MAP. The interested reader can refer to "Ego Development: A Full-Spectrum Theory of Vertical Growth and Meaning Making" (Cook-Greuter, 2014), which is highly recommended for a comprehensive overview of EDT.

for non-statisticians. The real issue, as will become clear, is the way statistical constructs have been misinterpreted in the STAGES validation article. I am not questioning any of the calculations that were done, I am making visible the lack of adequately connecting the statistical calculations to the very nature of the assessment technology being studied.[3]

Before proceeding, it is important to maintain a distinction between the *STAGES scoring system* on the one hand, and the *STAGES theoretical model* on the other. This present article is about the STAGES scoring system, not the STAGES theoretical model. I will, however, make some references to the model at the end, after I have examined the scoring system. I want to emphasize that I am not all against the underlying STAGES theory, and I suggest to the reader not to arrive at hasty conclusions about the theoretical model, only based on the examination of the scoring system.

Overview of Content

In this article, I will first define and contextualize the hypothesis being examined, before I give a general overview of the replication study. Next, I will explain the use of the Kappa statistic and why we must also take the scoring systems quality standards into account. When these items are explained, I will examine the actual data from the replication study. I will look at them from a few different angles, and then provide possible explanations for what the study reveals. After a discussion on what I see as implications of the analysis for positioning the STAGES narrative as an extension of ego development, I will show the long-term effect of too much scoring error. After the examination of the replication study, I will look briefly at the inter-rater study on the latest stages in the STAGES model (the MetAware Tier). I will discuss scoring at the MetAware Tier before making some comments about the STAGES theoretical model. The last sections are on the responses to STAGES before I wrap up with a conclusion.

Feel free to skim through sections if you are not too interested in the statistical explanations; you may come back to them later. The most important part of this article is the presentation and explanation of figures two and three, without which the subsequent discussions make less sense.

The Basic Hypothesis and Claim I Am Exploring

I will frame the hypothesis the STAGES validation article is testing this way: Can the STAGES theory, which is primarily based on three underlying parameters, be successfully implemented as a scoring system, based on three dimensions of language, such that it will replicate the MAP?

We read from the title of the STAGES article that their scoring system is validated by this study. But the claim goes further than this; the article states the following:

[3] All the work I've put into this article, including the writing, calculations, research, and dialogues with friends, I have done in my own free time. While I have done some professional work as a statistician for VeDa (Vertical Developmental Academy), first for Susanne Cook-Greuter and later Beena Sharma, that work has been done outside of my engagement with the STAGES model. No one asked me to do this investigation, and I have no economic interest in VeDa or STAGES.

> *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*. (p. 2)

What is important to recognize in this statement is 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. (The five objectives mentioned in the quote above are not relevant for this current article, the interested reader can go to the STAGES article.) What I want to emphasize is that this claim, as it directly relates to an existing developmental model, is a choice made by the authors of the article. It is very specific and has important consequences for the analysis and the discussion.

The second essential item in my article is the quality standards the scoring systems are based upon. Since the claim is that the valuable base of the MAP is retained, and the ego developmental model is strengthened, it must of course be proven that the inherent quality standards in the scoring system are retained. This will be explained below.

The Replication Study vs the Subsequent Analysis

In order to examine the validation claim properly, we need to make another important distinction, which is between the design and sampling methodology that provides the data, and the subsequent statistical analysis of the data once they are provided by the sampling technique. I will discuss each in the following two sections. I will refer to the study design, sampling technique and data gathering process as the study. This does not include the subsequent analysis which I view as a different topic.

Study Design and Sampling Technique

The study design and sampling conducted to provide the data (in order to then examine the STAGES scoring system) is well described in the validation article. In my opinion, it is straightforward and very well done. According to standard statistical principles one would select a set of inventories already scored by the MAP (the classical method used to assess ego developmental stages), and then rescore them using the STAGES method. There are four STAGES scorers who score independently of each other, and they have no direct information of the MAP assessment; they only see the sentence completions without any MAP comments. We have a total of 218 scorings, distributed over all stages (except the very latest ones). This turns out to be a sufficiently high number for my analysis and conclusions.

The data resulting from both the replication study from earlier stages and inter-rater study from later stages can be found here. (This link was provided in the STAGES validation article).

It is a simple spreadsheet that is well organized; all my subsequent examinations are done by using those data directly. So, thanks to the authors of the article the original data can be easily examined by anyone interested.

I will, in the following sections, provide several examples which should make the nature of the data from the replication study easy to understand.

Analysis of the STAGES Replication data

The statistical proof of the STAGES' hypothesis is summarized in table 2 in the STAGES article (page 13). I have copied table 2 and pasted it below for reference.

Table 2. Tier 1–2 Replicability of STAGES scores matching CG/L scores.

	N	Weighted Kappa	
		Stages 2.0 &2.5 separated	Stages 2.0 &2.5 combined
Scorer 1 (most experienced)	73*	0.95	0.94
Scorer 2	48	0.82	0.84
Scorer 3	48	0.79	0.81
Scorer 4	50	0.88	0.87
Mean (All)	**73**	**0.86**	**0.87**
Mean (excluding Scorer 1)	**73**	**0.83**	**0.84**

* All 73 inventories were scored by Scorer 1. Each inventory was scored by two of the other three scorers.

Table 2 presents what is called Kappa values, which we see range from 0.79 to 0.95. (The Kappa's are calculated for each of the 4 scorers, with and without a split between the two stages 2.0 and 2.5. I'll get back to this in the following). The Kappa values are numerical measures of how close two sets of ordinal data are to each other. Ordinal means numbers that indicate the exact position of something. The higher the Kappa, the better the match between the two data sets.

In order to make sense of these numbers, we must of course know what they actually mean. What is a Kappa of 0.95? Or a Kappa of 0.83? How should they be interpreted here? In order to do this the authors of the article write the following:

> *Kappa values can be interpreted as follows: κ < 0.0, no agreement; κ ¼ 0.0–0.20, slight agreement; κ ¼ 0.21–0.40, fair agreement; κ ¼ 0.41–0.60, moderate agreement; κ ¼ 0.61–0.80, substantial agreement; and κ ¼ 0.81–1.00, perfect agreement (Landis and Koch, 1977). We refer to the last category (κ ¼ 0.81–1.00) as "very strong" instead of the commonly used "perfect" agreement since Kappa ¼ 1.0 is the only value that indicates perfect (exact) agreement between the two sets of scores.* (p. 8)

So for the Kappa's in table 2 (except 0.79), using this interpretation the STAGES scorers all fall into the category of "very strong" agreement. From this the authors draw the conclusion that the STAGES scoring system is validated.

However, as I see it, there are three fundamental problems with the use of Kappa values as constituting a validation proof. I will go more deeply into these three issues in the next two sections. However, to summarize: First, one cannot use those interpretations quoted above

("substantial", "very strong" etc.) to directly validate a hypothesis. Second, the Kappa is not a good construct to use when we take into account how the two scoring systems relate to each other. 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.

The Interpretations of the Kappa Values

We must first understand that the linguistic interpretations ("moderate", "substantial" etc.) are descriptions referring to the mathematical realm. As such, they relate to the data merely as numbers, not to what the numbers *represent*. In this respect, the linguistic descriptions are more like guidelines, useful only under specific conditions. We cannot take a generalized term like "perfect" or "very strong" from a statistical construct like here and use it as a linguistic validation proof. When we do, we are relating to terms like "very strong" attached to a statistical number and overlook the very context from which the data themselves originate. Depending on the specifics of our hypothesis and the nature of the data under scrutiny, a Kappa value of 0.83 may be satisfactory, or it may not. A Kappa of 0.95 may be too low, or a Kappa of 0.79 may be high enough. This will make more sense a little later.

The second item about the usage of the Kappa is also related to the fact that the Kappa values only reflect the overall numerical relationship between two sets of ordinal data. In order to contextualize that numerical relationship properly, we need to know about any prior relationship between the two data sets. The article explains this well:

> *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.* (p. 11)

This means that the STAGES scoring system is directly derived from the MAP. The STAGES scoring system is a *wholly derivative method*, and a high correlation, or correspondence, is already a given. In light of O'Fallon's work, anything but a high Kappa score would be very surprising. If STAGES was a new model conceived from outside of ego developmental theory, the "very strong" labeling of the Kappa's could be more relevant, numerically speaking. But even if that was the case, we are still not taking into account the quality standards the two scoring systems are based upon.

Which leads us to the third, and most important, point: The quality standard of the scoring systems, and how that relates to Kappa.

The Quality Standards of the Two Scoring Systems

As the STAGES article explains:

> *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.* (p. 6)

We see here that a minimum of 85% correct matching is the quality standard used by the STAGES scoring system. In other words, when a STAGES scorer is hired, we should expect that he or she has at least an 85% chance of giving the correct score on our assessment. Correct here means in accordance with the postulated underlying theory itself. It turns out the 85% standard is the same for training scorers for the MAP (Cook-Greuter, personal communication). Keep in mind that this is the *minimum* requirement, so in reality it will be equal or higher.

There is also a secondary aspect of the quality standard I must mention. The maximum number of mismatches, which is 100% - 85% = 15%, should not be more than one stage apart from the correct score. It is not explicitly mentioned in the article, but I must assume it is also relevant for STAGES. I have this from a communication with Cook-Greuter, where she said that when someone during the exam misses by more than one stage, for the inventory as a whole, "*it is a serious concern*". This will be discussed after I have presented more data.

In the next sections I will present some figures. The Kappa's are in each case a single summary function of all the positions we see in these figures. If there is a direct match, that particular score gives maximal contribution to the Kappa value. Depending on the number of categories, a deviation of one stage gives a little less contribution, two stage deviation a little less than that and so on. The final Kappa score is the sum of all these individual scores.

As I will show, a fundamental limitation of using the Kappa here, is that it does not properly account for 2 stage deviations, or for more than 15% one stage deviations. As a result, in this case Kappa is simply concealing that the STAGES scoring systems own quality standard has not been met.

Examining the Data from the Replication Study

What I did first was to check the data against the basic quality standards by which the scoring systems are defined. The easiest way to start is simply by arranging the data in a cross table and calculating the number of correct matches. One version of such a table is already presented in an appendix to the STAGES validation article, table 5. I have copied and pasted table 5 below.

The stage labeling in the row above the frame is from both systems, the STAGES notation and MAP in parenthesis. We see for example that stage 1.5 in STAGES is called 2/3 in EDT. Both systems have good reasons for these numbers, but for simplicity I will only use the STAGES labels in the following. I will also use the Wilberian notions of colors as markers of developmental altitude: The STAGES numbers correlate to developmental altitudes roughly in this way: 1.0 Magenta, 1.5 Red, 2.0 Red/Amber, 2.5 Amber, 3.0 Amber/Orange, 3.5 Orange, 4.0 Green and 4.5 Teal. (See Wilber, 2017 for a more detailed description).

Table 5 from the article.

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.)

Explanation of Table 5

In this table we see all of the data from the replication study inside the frame. Outside the frame, vertically and horizontally, are summaries of the number of entries for each stage ("Total"), with percentages of "Agrmnt" (direct match) and "Agree +/- 1" (one stage deviation) for each of the stages.

The MAP stages are presented vertically, while the STAGES' stages horizontally. We have, for example, the first vertical column showing 12 STAGES scorings on inventories scored at 1.0 by the MAP. Out of these we have 4 scored as 1.0 by STAGES, 5 scored as 1.5 and 3 scored as 2.0-2.5. There were 4 inventories at 1.0 (by the MAP), O'Fallon scored all of them, while 2 of the other 3 scorers (independently chosen in each case) scored each as well. So we have 4 inventories X 3 independent STAGES scorings = 12. The number of correct matches for each stage are in bold and in red font, in the diagonal from the upper left corner down towards the lower right corner.

How well STAGES replicate the MAP, for each stage, will be found in the horizontal row under the frame labeled "Agrmnt". The horizontal row "Agree +/- 1" shows the amount of 2+ stage deviations (the gap between 100% and the percentages listed).

The vertical columns outside of the frame to the right, with the same labels, have less interpretive value and will not be commented upon. The reason is that the sample size for each stage is different from the relative size of each stage in any actual population, an important point I will come back to in the section *The long-term consequence of too much scoring error*.

Interpretation of Table 5

In the table we see that the percentages of correct matches for each stage found in the horizontal lines under the table ("Agrmnt") goes from 33.3% to 77.8%. For the overall matching for all the inventories, we can simply count the correct matches (in bold) and divide this sum by the total number, which gives 135/218 = 61.9%. We thus have that the overall correspondence 61.9% is far

away from 85%, and that this quality standard is not met for any single stage. From table 2, we find that the mean weighted Kappa for all the 4 scorers (and 2.0 and 2.5 combined) is 0.87, a number which is derived from the very same data as in this table 5. We see that a Kappa value of 0.87 has nothing directly to do with the percentage of correct matching, and that the concurrent label of "very strong" agreement cannot alone be used as a validation criterion.

The correspondence between the two scoring systems may seem visually satisfactory to the eye looking at this table, but that is of course not proof of validation. It merely displays the a priori expected close relationship between the two scoring systems, when we know how the STAGES scoring system is derived from the MAP. This table appeared sound to me when I first saw it, but that was before I was aware of the quality standards and how the scoring systems actually work.

For the two earliest stages 1.0 and 1.5, the overall matches are 33.3% and 41.7%. The STAGES validation article says about them that:

> *it would seem unacceptably low…in real applications less than 1-2% of adult individuals are expected to score at these levels so practically the mismatch is not very important.* (p. 14)

Applying scientific rigor we see that the STAGES scoring system is not *fully* validated as an alternative to the MAP, by the admission of the authors themselves.

One can use table 5 alone to prove that the replication claim cannot be true, a proof which I presented to the authors of the validation article in October 2020. However, to make my observations easier to understand and debate, I will first separate the STAGES' master scorer Terri O'Fallon from the other 3 scorers. This will provide more clarity on what is going on.

A Display of O'Fallon's Scores

As the STAGES validation article states:

> *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.* (p. 8)

It is a quality sign that this special precaution is recognized, and we also see this done with the Kappa's in table 2. However, there is another important reason for this precaution: We read in the article that an experienced scorer of the MAP:

> *have memorized the gist of most categories and can score most inventories without consulting the manual.* (p. 3)

As the article further points out:

O'Fallon trained as a scorer in the CG/L method and did scoring and related consulting for about six years. (p. 11)

And that she was exposed to:

well over a thousand CG/L inventories. (p. 11)

So the situation is that O'Fallon already knows how the MAP works and could score most inventories without consulting the manual (used for the MAP). The article says that O'Fallon:

was the only STAGES scorer who had also studied the CG/L method. (p. 8)

and that

all four scorers were familiar with developmental models prior to learning how to score. (p. 7)

So the other 3 scorers were familiar with developmental models but had not studied the MAP scoring protocols. Figure 1 is a version of the STAGES' table 5, with only O'Fallon own scores.

		MAP Scores								
		1.0(2)	1.5 (2/3)	2.0 - 2.5 (D3-3)	3.0 (3/4)	3.5 (4)	4.0 (4/5)	4.5 (5)	Total	Agreement
Terri O'Fallon	1.0	**2**							2	100%
	1.5	1	**3**	1					5	60%
	2.0-2.5	1	1	**14**					16	88%
	3.0			1	**11**				12	92%
	3.5			1	1	**11**	1	1	15	73%
	4.0					1	**9**	2	12	75%
	4.5						2	**9**	11	82%
	Total	4	4	17	12	12	12	12	73	
	Agreement	50%	75%	82%	92%	92%	75%	75%		**80.8%**

Figure 1. STAGES table 5 with only O'Fallon scores.

We see that the overall agreement is now 80.8% (right bottom corner) and quite close to 85%, which is the minimum agreement acceptable to be certified as a MAP or STAGES scorer. So O'Fallon is fairly close to "passing the test" directly so to speak. We also see that two of the correspondences on separate stages are above 85% (bottom row "Agreement").

Figure 1 illustrates well what is meant by the 85% minimum direct match criterion, which is so fundamental to the quality of the MAP assessment. It all has to do with *necessary precision* in the measuring instrument.

The corresponding Kappa for the data in figure 1is found to be 0.94 (see STAGES' table 2). So we see again that for our two particular sets of ordinal data under investigation (MAP vs STAGES), a Kappa must be higher than 0.81. The Kappa limit at 0.81 and the corresponding term "very close" agreement cannot be used here to justify any validation claim. In fact, even a Kappa as high as 0.94 has not met the minimum 85% direct match criterion.

In the case with figure 1, we see three instances of a mismatch of 2 stages. I would imagine that in a practical situation, these 3 inventories would be examined and discussed with the trainee, before the certification process could proceed to its possible conclusion.

Looking at the 3 Scorers Not Familiar With the MAP

Now I will turn the focus on the 3 certified STAGES scorers not familiar with the MAP. As the article recognizes, special precautions must be made to make calculations without O'Fallon's own scores. Since the 3 scorers are not familiar with the MAP, note that it is with this data set we can *truly examine the STAGES scoring system in comparison to the original MAP*. We are now approaching the crux of the matter.

Removing O' Fallon's own scores we now have a proper unbiased situation, statistically speaking. Since the 3 certified STAGES scorers are not familiar with the MAP, we are now directly looking at the STAGES scoring system's capacity to replicate the MAP. We have removed from our data any outside information that could influence the results. The results for the 3 scorers are presented in figure 2.

		MAP Scoring								
		1.0 (2)	1.5 (2/3)	2.0 - 2.5 (D3-3)	3.0 (3/4)	3.5 (4)	4.0 (4/5)	4.5 (5)	Total	Agreement
3 INDEPENDENT SCORERS	1.0	**2**							2	100%
	1.5	4	**2**	2	2				10	20%
	2.0-2.5	2	4	**22**	9	1			38	58%
	3.0		2	8	**10**	1			21	48%
	3.5				3	**17**	6	2	28	61%
	4.0					2	**10**	4	16	63%
	4.5			2		3	7	**14**	26	54%
	5.0						1	3	4	0%
	5.5							1		0%
	Total	8	8	34	24	24	24	24	146	
	Agreement	25.0%	25.0%	64.7%	41.7%	70.8%	41.7%	58.3%		**52.7%**

Figure 2. STAGES scoring without O'Fallon.

At this point I invite the reader to pause and take in what figure 2 tells us. We are here looking directly at the STAGES scoring systems capacity to replicate the MAP; we are looking at what hides behind the Kappa score.

A Kappa based on these 146 data points is 0.84 (STAGES' table 2), within the "very strong" agreement category. The reason for this relatively high Kappa number is that the data are clustered around the diagonal from the upper left to the lower right, as already expected. But in light of the actual quality standard, adopted by STAGES, the number of correct matches is only 52.7%. Out of the 7 stages measured here, none are close to 85%, and 4 have a direct match below 50%. The STAGES capacity to measure Magenta and Red accurately are both 25%. For Amber/Orange and Green the accuracy is 41.7%. And these numbers are, as we saw above, recognized as too low by the authors of the verification article themselves.

Further, we find 16 mismatches over 2 or more stages, which is 11.0 %. And we have not yet separated the two stages 2.0 and 2.5. This separation is left to the appendix, as it requires a bit more explanation. The number of correct matches drops below 50%, and the 2+ stage deviations go up to 18.5%.

What we see is that the STAGES scoring system does not work as a way to replicate the MAP. And it is not that it "almost works" or is "not quite close enough". It simply fails to measure ego developmental stages with any kind of necessary precision. We only needed to look behind the Kappa values, take O'Fallon's scores out (as the article points out is necessary) and take into account the STAGES' own quality standard. It is this simple. And it is revealed by STAGES' own study.

We now see this statement in a different light:

> *STAGES retains the valuable base of its predecessors…… and strengthen the ego development model and its assessment.* (p. 2)

This statement can only be made if we believe we have proof of validation. But instead of proof of validation, we have another proof. We have solid proof that the STAGES scoring system does *not* replicate the MAP. The probability of getting figure 2 by coincidence, *if the 85% direct match criterion was met*, can be shown to be no more than 0.000000701% (see Appendix). Which is a very small number, and a solid a proof as you need.

Additional Statistical Calculations

Some may ask at this point if there are any alternative statistical calculations that will yield a different result? My answer is no. The direct presentation of figure 2 remains the same regardless and is the most basic way to display the data. The statistician has many tools in his toolkit, but when you have solid proof by using a simple technique, you go no further. See also the appendix *The Role of the Statisticians*.

Anybody is of course free to check my calculations; the data is available to everyone and making figure 2 can be done by anyone who can handle an Excel spreadsheet. But you cannot possibly come to another conclusion, unless you disregard the quality standard the scoring systems are defined by. Either way, the basic replication statement in the STAGES article is refuted.

Once this basic fact of non-replication is established, there are other calculations that can be made in order to shed more light on the situation. I will present some below, after more comments on the results revealed by figure 2.

Discussing the Results from Figure 2

Reading this may be disconcerting to some. But I hope to avoid attempts to question the basic result the data actually shows. I would strongly object to that; the sampling technique itself is very well done. Keeping the distinction between the sampling technique and the analysis, the issue is only the part of the analysis that is the misinterpretation of the Kappa values.

Neither can we explain this away by saying that the scorers are better equipped now than before. At the time of the replication study, the scorers did not have any written material they could use in the scoring. Today they have (see more on this in the section *Discussion on why the independent scorers do not score accurately enough* below), but that is not relevant for our discussion on the replication study. We cannot jettison the result of the replication study until we have another study conclusively demonstrating otherwise.

A new study can go both ways, as the STAGES rebuttal (see appendix) confirms (according to the rebuttal the replication study, not published at the time, would reveal a strong proof of the validation of the STAGES scoring system):

it is always possible that future research will not show results as strong as the recent replicability study.

The measured discrepancies are very, very far away from the quality standard. This is the empirical truth about the STAGES scoring system as it relates to ego developmental stages, and it will stay that way until a possible future similar study would conclusively prove otherwise. See also the section below *Discussion on why the independent scorers do not score accurately enough*.

Questioning the results presented here would also mean hampering the possibility of improving the STAGES scoring system or inventing a new one. This original validation study is methodologically solid, important and the results are clear and in fact very informative. From a scientific viewpoint, there is nothing inherently "negative" (or "positive") about the study and the results in and of itself. It is only "negative" if you are personally, emotionally and/or financially invested in the results being different.

I will also caution against referring to different STAGES studies before the results from the replication study are fully accepted. There may be other types of studies done showing benefits of other aspects of STAGES, but none of those could make any difference to the replication study.

I welcome counterarguments on my assumptions, calculations, and comments here. Anyone is free to do that, but I ask you first to read the rest of this article, reflect on the totality and stay on the topic.

Comments on the Analysis

There are many questions that may arise now: First, why has this simple truth about the STAGES scoring system apparently gone under the radar by the authors of the article? Second, why do the 3 trained STAGES scorers miss the mark to this extent, so far below the quality the STAGES scoring system itself is defined by? Third, what are the implications of what is revealed here; does it really matter that much? Fourth, what about the underlying theory itself? And why hasn't this been spotted by people engaged with STAGES? I will discuss these questions, as I see them, in the rest of this article.

Possible Explanations for the Misinterpretation of the Kappa's

The first question can be better understood by realizing that 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. See also the appendix in the chapter on *The Role of the Statisticians*.

On the other side, it appears that the non-statisticians of the article have relied solely on the linguistic interpretations of the Kappa values, and not taken the time to properly investigate the actual deviations between the two scoring systems, or adequately question the statistical analysis.

And why can the 3 STAGES scorers not detect this themselves? As quoted above, O'Fallon

was the only STAGES scorer who had also studied the CG/L method. (p. 8)

My understanding is that any STAGES scorer without enough experiential understanding of ego developmental theory and the concurrent markers of each stage will simply have less chance of detecting mismatches.

Before I discuss other implications of the results in figure 2, I will check for the possible difference between ego development and STAGES. Since the basic truth of non-replication is established, I can look for possible explanations.

Does the STAGES Scoring System Measure Something Different than Ego Developmental Stages?

A hypothesis worth considering is whether the STAGES scoring system *systematically* measures something different than ego developmental stages. 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. Enactment is a fundamental Integral principle which I will not fully elaborate upon here, but from this perspective, there is no "ego development" lying around waiting to be discovered, it is enacted and brought forth by the particular methodology applied by Loevinger and expanded upon by Cook-Greuter. In Ken Wilber's *Religion of Tomorrow*, (2017) we read that:

> Thus, an Integral Psychograph, besides presenting some of the most common basic parameters found in most models (such as altitude, levels, lines, states, types and so forth), will on different occasions include Piaget's stages, or Kurt Fisher's stages, or Jane Loevinger's stages, or Lawrence Kohlberg's stages, or Clare Graves stages, or Terri O'Fallon's stages, since something specifically unique and useful can be found in every one of those models – each of them is presenting a unique perspective on the altitude and specifics of development in a particular line or group of lines – and this openness and inclusion is clearly something that none of those models, on their own will do… (p. 512)

Using Wilber's integral framework here, something unique and useful can be found in each of these developmental models. He explicitly mentions the two we are investigating here, ego development (Loevinger) and STAGES (O'Fallon). Each of these two present a unique perspective on development, alongside the other important developmental constructs. Each one is true but partial.

Comparing the STAGES Scorers with the STAGES Master Scorer

Based on the comments above, we can *hypothesize* that a person could have a center of gravity according to ego developmental theory, and a different one in terms of STAGES' description of stages. This could explain the results in figure 2 and be in accordance with Integral Theory and the Wilber quote above. We can check if the STAGES scoring system captures this in the following way: Instead of comparing the 3 STAGES scorers with the result from the MAP, we can compare them to the result of the master STAGES scorer, O'Fallon herself. Note that we are now totally within the STAGES theory and scoring system itself, as we are not comparing it with the MAP or ego development at all. We can split 2.0 and 2.5 without further explanations, as they are now treated equally by all the scorers. We then get figure 3.

		Master scorer (Terri O'Fallon)								TOTAL	Agreement:
		1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5		
3 INDEPENDENT STAGES SCORERS	1.0	**1**	1							2	50.0%
	1.5	3	**4**	1		2				10	40.0%
	2.0		3	**23**		8				34	67.6%
	2.5		1	1	**0**	1	1			4	0.0%
	3.0		1	5	2	**11**	2			21	52.4%
	3.5					2	**20**	6		28	71.4%
	4.0						2	**12**	2	16	75.0%
	4.5						5	5	**16**	26	61.5%
	5.0							1	3	4	
	5.5								1	1	
TOTAL:		4	10	30	2	24	30	24	22	146	
Agreement:		25.0%	40.0%	76.7%	0.0%	45.8%	66.7%	50.0%	72.7%		**59.6%**

Figure 3. Comparison of the STAGES scorers and the master scorer.

The overall agreement is now 59.6%. For each stage individually, we have from 0.0% to 76.7% direct matching. We are still far away from the 85% criteria, which is not even reached for any single one of the individual stages. And there are 25 mismatches over 2 or more stages, which is 17.1%.

Note that we are now looking at scoring within the STAGES own system, not as it attempts to assess ego developmental stages. What the article explicitly states about the certification process of STAGES scorers, requiring 85% direct matching to the master scorer, does not hold at all in this real-life situation.

This brings me to the next point, a closer look at the STAGES certification process. I don't know exactly how it is done, so I have to rely on the descriptions in the STAGES validation article. But first I will check one last possibility.

Comparing the 3 Scorers to Each Other

Another possibility, at least in theory, is that the STAGES scorers not familiar with the MAP do something systematically different than O'Fallon since they are equal in the sense that they do not know the MAP protocols.

If we compare the 3 scorers individually, 2 vs 3, 2 vs 4 and 3 vs 4, we will find that they have the same score in 63.9%, 58.3% and 56.5% of the inventories, respectively. The amount of 2+ mismatching is 13.6%, 8.3% and 17.4%. (This is simply a matter of counting the matches in each case, using the aforementioned spreadsheet. These 3 scorers are labeled scorer 2, scorer 3 and scorer 4 in this spreadsheet.)

So the 3 independent scorers do not agree with the MAP, nor with the STAGES master scorer, and they do not agree between themselves. What I mean by "not agreeing" is gauged against the amount of *scoring precision* the STAGES scoring system itself claims it has.

Statistically speaking, there is simply too much random noise within the STAGES scoring system. Random noise here means scoring error and refers to any outside influence that is not related to the nature of the developmental stages themselves. I will come back to why I think this is the case after a discussion on the certification process.

The STAGES Certification Process vs the Replication Study

The article says about O'Fallon's initial work to define the STAGES scoring system, as it is derived from the MAP (using the manual):

> *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.* (p. 11)

But STAGES' own study shows that the 3 scorers do not score accurately. Remember that we see a correspondence far below the stated quality standard. Why this happens, I cannot say for sure. These comments from the STAGES article may give an important hint however:

> *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.* (p. 9)

I understand this to mean that the 85% correct score is based on only 10 inventories. It looks as if the certification process can go on *until this criterion is achieved.* But these 10 last inventories will not be representative for the overall long-term reliability, which is what the replication study measures. The STAGES article further states that:

> *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).* (p. 9)

I interpret this as saying that we are looking at the final 10 pre-certification scores, but we already know they have at least 85% correct score *before* examining them. This would then be a biased sample of inventories and doesn't say much about the overall long-term scoring accuracy.

In any event, the issue is that the information provided on the certification process does not say anything conclusive about how well the scoring system functions long term, outside of the certification environment.

This shows how important the replication study is. That study is not about results from the scoring exam; it is about what is happening in real life. To see how well something functions in real life is exactly why we do these kinds of statistical studies. Generally speaking, what we see during training, or in the laboratory, may not be what we see in real life.

The 3 independent scorers were taken outside of the training environment, without any feedback or supervision. There are as many as 146 scorings in the sample behind figure 2 and figure 3. The sampling technique is very well done, and the data is reliable.

Discussion on Why the Independent Scorers do not Score Accurately Enough

We see that these certified STAGES scorers, with only the three questions to guide them, and no prior exposure to EDT or how the MAP works, produce figure 2. Terri O'Fallon on the other hand, is well trained in EDT and the MAP and therefore has an acquired direct intuitive knowing for which category any sentence should be rated at. We see this in figure 1.

What I conclude is that O'Fallon's acquired MAP-competency cannot fully be transferred into the three questions. The replication study verifies that the three questions are not enough for this scoring system to work adequately.

And it is not that it "kind of" does not work, it just doesn't work at all. It is not that we see 75% or even 65% correct matching, which would itself not be satisfactory for replication, we are below 50%! The evidence is overwhelming. It is not my personal opinion, it is a verified fact revealed by a solid research study, done by the STAGES community itself.

The article concludes this about the STAGES scoring system in relation to the MAP:

> *The theoretical principles, once learned, eliminate the need for a manual of sentence completions...* (p. 10)

But that is a statement which can only be made if we believe the empirical research constitutes a validation proof. It does not, which clearly indicated that we cannot eliminate the need for the manual of sentence completions this way.

Today, the STAGES scorers use their own STAGES manual with details about the scoring of each parameter, and what factors to consider in aggregating completions into the final score (Tom Murray, personal conversation). At the time of the replication study, they did not have any written material to support them. On the other hand, the MAP scorers have descriptions of 14 criteria they can use to score correctly, in addition to the manual of sentence completions. And of course, scorers of both systems have gone through their respective certification training.

In an attempt to explain the results of the replication study I need to examine what are the differences between the two approaches to scoring. I know from the article that none of the 3 independent scorers have been trained in ego developmental theory, only the STAGES model. Further, the MAP scorers have a manual of sentence completions, while the STAGES scorers do not. The MAP manual has examples of sentence completions for all the 36 standard questions the MAP is based on, for all the EDT stages. That is, whenever there is a direct match in an inventory found in the MAP manual, the scorer directly knows which stage the sentence should be scored at. We can say that the MAP manual this way implicitly *contains within it ego developmental theory,* not as a description of the theory, but how it is embedded in sentence stems exemplars. The independent STAGES scorers, on the other hand, have no exposure to the actual EDT, nor any manual of sentence completions; they have to rely on their own understanding of the 3 questions (and today the support of a STAGES manual). These two main differences (knowledge of ETD and a manual of sentence completions) inevitably open up to more scoring variability and thus errors coming from the STAGES scorers, than with the MAP scorers. Scoring error is any factor influencing the scoring which is not related to the real nature of the ego developmental stages themselves.

I believe this is the main reason why the STAGES scoring system does not work as a way to assess ego developmental stages. We can only "eliminate" the manual of sentence completions if the scorer has so much training with it that they have learned to internalize it. Which means it is not eliminated, it is internalized.

And I cannot see how any deeper experience as a STAGES scorer, or a written STAGES manual, can alleviate this fundamental issue; the measured discrepancies between the two scoring systems are simply too high. Again, until a new study conclusively demonstrates otherwise, this is the empirical truth we have. We can also understand this better by looking a little closer inside the scoring systems, see the Appendix. In the next section I will give some comments on improving the STAGES scoring system.

Improving the STAGES Scoring System

Based on the discussion above, there are two things I think need to happen if a STAGES scoring system was to work adequately.

First, 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, especially since the MAP is coming from a different developmental construct. This implicit reliance is there today, to some extent, since O'Fallon relied upon the MAP manual of sentence completion at the time she was

drafting and refining general scoring rules for the STAGES scoring system. I believe this can explain much of what has been revealed by the STAGES replication study, where we have also seen that the trained scorers cannot replicate O'Fallon's own scores, which must be influenced by her prior exposure to the MAP definitions.

Second, the 36 sentence stems and that particular assessment system may not be the best instrument. This particular method has co-evolved with ego development over 50+ years and the two, EDT and the MAP, are intimately connected.

I am not sure what a different STAGES scoring system would look like. I know the Lectical approach to scoring, which is completely different from the MAP. Why not invent a new and novel one for STAGES, instead of copying the MAP? If the STAGES scoring system then proves to replicate its own developmental stages with the stated quality standard, by a new study as solid as the replication study, I will be very happy to endorse it publicly.

Implications for the STAGES Narrative

Before this particular discussion, remember that in this article I use the term "ego developmental theory" and "EDT" interchangeably to refer to the L/C-G system, as pointed out in footnote 2. This is a narrow definition and not the only valid use of the term. But since "ego developmental theory" and EDT is generally known to refer to the L/C-G model, and I need to make a clear distinction between the two soring systems, I choose this definition for clarity. Outside of this particular context, I have no problem at all if the STAGES theoretical model qualifies to be another version of "ego development" in a broader definition of the term. From this perspective I see any discussion here as purely semantical. From another perspective, I see STAGES using the term "ego development" as problematic when the rational is that STAGES "strengthens the ego developmental model and its assessments" referring to the L/C-G system and the replication study.

A consequence of the replication study is that since the foundational validation claim in the article is seen not to hold up under scrutiny, many other claims about STAGES are automatically questioned too. STAGES could have defined itself as an independent new theory with its own scoring system and left to future application and research to answer the question as to its place alongside other developmental theories and measuring techniques. That is not the case, as the STAGES article states very explicitly.

For example, the article states this about 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.* (p. 10)

And we have this statement:

> *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…* (p. 9)

The validation article has more statements aimed at building a case for all the benefits of the STAGES theoretical model and its scoring system, as an improvement of ego developmental theory and the MAP. But when the basic validation proof collapses, so does *the validity of any statement based on the truth of the validation claim*. The two statements above are both invalidated by STAGES' own research.

We see the larger implications of the fact that STAGES has identified (and marketed) itself as an improvement of ego developmental theory, including its MAP assessment. STAGES is *derived from* ego developmental theory, which is a legitimate thing to do in the first place. But STAGES postulates itself as retaining its predecessors base and updating that model and the assessments. It incorrectly uses validation studies done on the MAP to justify its own theory and scoring method.

This does not make ego development and the MAP "100% true". EDT is a developmental construct with a long history which has proven to be very useful in many contexts. It is subject to continual refinements from those working with it professionally. And there is always the importance of not taking its potential usefulness too far and out of the proper context.

I am a defender of rigor and intellectual honesty when it comes to how we relate to empirical evidence. That said, I want to emphasize that I am not all against the STAGES theoretical model. I am among those who see useful *aspects of* it, but the belief that three underlying parameters can explain and replace the whole of ego developmental theory and its assessment is simply not true. The narrative of STAGES as being about ego development only confuses the important distinctions between the two theories.

With my best goodwill towards STAGES, I cannot see why this narrative is necessary and I cannot see how it can possibly benefit STAGES in the long run. The results from the replication study should make this very difficult to ignore and thus easy to appreciate. A clean differentiation of STAGES from EDT will make it much easier to continue the necessary dialogue, and research, around the nature of development in general. There is much more to be said on the important distinctions between EDT and STAGES, but that is beyond the scope of this article.

The next section will show the long-term consequences of too much scoring error. We will see the crucial importance of scoring precision.

The Long-term Consequence of too much Scoring Error

Over time, the effect of the lack of adequate scoring precision will become apparent in any population being scored. A comment in the STAGES verification article relating to table 5 is worth examining in this respect:

> *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 Strategist individuals as Pluralist (or lower). Overall, there does not appear to be a significant shift, higher or lower, in STAGES vs. CG/L scoring. (p. 14)

In this section I will focus on the results from the replication study, but I will return to this concern again when I discuss the MetAware Tier.

One problem with the quote above is 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.

Using 4.5 Teal as the example, we have that in most real-life populations, there will be many more people at 3.5 (Orange) than at 4.0 (Green), and more at 4.0 (Green) than 4.5 (Teal). A high relative prevalence of Orange and Green as compared to Teal, in combination with too high scoring error, will give a high relative presence of people scored *as* Teal, while they are in fact *at* Orange or Green. According to Wilber's general altitude descriptions, Teal altitude would be the entry into 2nd Tier consciousness.

So, in order to check the concern above I will combine the results from the replication study (figure 2) with some real-life populations. I will use 4 different population distributions I found in table 3 presented in the book chapter *The Aware Leader: Supporting Post-Autonomous Leadership Development* (Lynam et al., 2020). I have pasted it below.

Table 3. Real life stage distributions

Table 3. Comparison of Percentage Stage Distribution in Six Different Samples (Anderson, 2014; Cook-Greuter, 2013; O'Fallon et al., 2018).

	535 Leaders, UK, Pre-2000 Torbert and Cook-Greuter	497 Leaders USA, Pre-2000 Torbert and Cook-Greuter	4,510 Mixed Adult Population, USA, 1999 Cook-Greuter	3,397 International Self-selected, 2000–2007 Cook-Greuter	905 International, Self-selected, 2014–2018 O'Fallon	150 GTC Pre-test Data 2011–2019
Conformist	1.7	8.2	4.3	72		
Expert	21.1	47.8	11.3			
Achiever	33.5	34.8	36.5		18	9
Pluralist	23.4	5.0	11.3	15.5	20	18
Strategist	13.5	1.4	4.9	9.0	33	41
Construct Aware	5.6	<1	1.5	3.0	18	22
Transpersonal	0.09	<1	0.05	0.5	10	9
Universal					0.09	
Illumined						

The 4 distributions to the left are the results of MAP scoring, the 2 to the right are results of STAGES scoring (I'll come back to these last 2 below discussing the MetAware Tier).

Figure 2 provides estimated probabilities of scoring someone *as* 4.5 Teal by STAGES, while they can be *at* 3.5 Orange, *at* 4.0 Green or *at* 4.5 Teal. And table 3 provides the *relative size* of

Orange, Green and Teal in 4 different populations. The calculations are statistically basic and straightforward to do and are described step by step in the appendix. Performing these calculations, we arrive at figure 4:

	Distribution of persons scored as 4.5 by STAGES			
MAP stage	Pop 1	Pop 2	Pop 3	Pop 4
3.5	22.2%	65.6%	42.6%	27.7%
4.0	36.2%	22.0%	30.8%	33.5%
4.5	41.7%	12.3%	26.7%	38.8%

Figure 4. Scoring distribution for stage 4.5 in STAGES.

Figure 4 shows that for the 4 real-life populations we have, the actual number of persons scored *as* 4.5 by STAGES, *actually being at* 4.5, Teal altitude, 2nd Tier consciousness, will vary from 12.3% to 41.7%. In each case, they will be a minority with the majority at an earlier stage. In other words, the anecdotal concerns by colleagues of giving participants a false sense of higher development are real concerns.

What is labeled Pop 3 in figure 4 (taken from the pasted table 3) is the mix adult population of 4,510 inventories scored by the MAP. This one is often referred to as the best distribution estimate we have of ego developmental stages in the actual population (there are also a few referrals to numbers from this particular population in the STAGES validation article). Using the STAGES scoring system on this population and focusing on those scored as 4.5, we will over time approximate the percentage distribution vertically presented in figure 4 under the label "Pop 3". The likelihood of a randomly chosen person actually being at 4.5 is only 26.7%, and the chance is *greater* (42.6%) that the person scored is actually at 3.5.

This may sound strange to the reader at first glance. However, the calculations are straightforward and there is nothing speculative about them; they are actually standard calculations in diagnostic situations like this.

We can now appreciate why Cook-Greuter says it is a *serious concern* with deviations over more than one stage. There is, from an ego developmental perspective, a very big difference between a person at 3.5 (Orange altitude) and one at 4.5 (Teal altitude). To score someone who is at Orange as Teal is not a good idea for the Orange client. To earn the relatively rare stabilization of the Teal altitude, everyone needs to make the fundamental (and sometimes difficult) shift from Orange to Green first. Then Green has to be fleshed out and lived through before the Teal altitude can take hold. It is not uncommon for Orange, who has read Integral, to talk Teal, as a person with sufficient intelligence at formal operational cognition (Orange) can learn virtually any theory in an objective fashion. And this is not a "mistake" made by Orange, is it simply how Orange usually makes meaning of stage-development.

As for any group of people scored as 4.5 Teal by STAGES, from any one of these 4 real-life populations, they will in reality consist of Orange, Green and Teal as we see in figure 4 above. This means that any such "Teal" group will not have the group center of gravity at 2nd Tier, but at 1st Tier. Yet they will all believe they are Teal and 2nd Tier since that is what the scoring says. If

they are given any problem to work on collectively, any solution will be an inevitable result of their actual developmental strength, not what the scoring says. So when we look at groups this way, from this particular developmental lens, we see that too much scoring error can have some unfortunate consequences.

This example shows that the lack of scoring precision will, over time, given the estimated probabilities in figure 2, contribute to a misunderstanding of what constitutes the integral notion of 2nd Tier consciousness. But the scoring error goes fo7r all the stages, so this will result in an increased confusion about development in general. This is the inevitable long-term consequence of too much scoring error.

Next, I will look at the MetAware Tier, before a discussion on the positive responses to STAGES, the theoretical model and then a final conclusion.

The MetAware Tier

There is also the inter-rater study for the 4 later stages (5.0 - 6.5) discussed in the article, and there are more Kappa values presented as proof of high inter-rater correspondence. I will not discuss those Kappa's, but simply make the figure for the stages 5.0 - 6.5 the same way I did for the replication study in figure 3. This gives figure 5.

Note that EDT (and MAP) has two post-autonomous stages representing the same territory, developmentally speaking. Since STAGES defines this territory differently, however, a deeper discussion of the two theories is beyond the scope of this article and I will not go into this at any length here. I have a good basic understanding of the models, but I'm not qualified to go into this at a level of too many theoretical details.

		Master scorer (Terri O'Fallon)				TOTAL	Agreement:
		5.0	5.5	6.0	6.5		
3 INDEPENDENT STAGES SCORERS	4.0	1				1	0.0%
	4.5	12	3			15	0.0%
	5.0	**21**	7			28	75.0%
	5.5	12	**43**	7	2	64	67.2%
	6.0		3	**10**	1	14	71.4%
	6.5			1	**7**	8	87.5%
TOTAL:		46	56	18	10	130	
Agreement:		45.7%	76.8%	55.6%	70.0%		**62.3%**

Figure 5. Scoring distribution for stages 5.0 – 6.5 in STAGES.

The overall agreement in figure 5 is 62.3%, only slightly higher than 59.4% as for the earlier levels in figure 3. So we observe the same phenomena for these later stages as the earlier ones, as

it relates to how the 3 scorers match the master scorer. We are far away from 85% here too, which is still not reached for any of the stages individually.

Note that the "existence" of these later stages as verified developmental stages (as they are defined by STAGES) is another discussion. 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. So even though the results in figure 5 were different, it would only prove that the scorers did the same thing more consistently. In one of my dialogues with Nayak Polissar (the lead statistician behind the replication and this inter-rater study) we both agreed that the inter-rater study carries no empirical evidence of the "real existence" of these late stages (Nayak Polissar, personal conversation). Conversely, figure 5 is not proof that these stages do not "exist", and certainly not that hypothesizing about them is not a worthwhile endeavor.

I am very positively oriented towards building new models about higher developmental stages, aiming at deepening our understanding of various aspects of this territory. I see the usefulness of having more than one developmental model, especially the STAGES theoretical model in conjunction with other models. In that respect, I am another supporter of the STAGES endeavor to shed more light on whatever happens when development continues into the higher realms of what may be developmentally possible for humanity. I have respect for all the work the STAGES community have done in this area.

But I am also a supporter of rigorous application of statistics, and I am aware of our capacities for self-deception when it comes to projecting ourselves into these higher stages. Again, to keep the distinction between the STAGES scoring system and the STAGES theoretical model is crucial. In the next section I will discuss scoring at the MetAware Tier.

Scoring at the MetAware Tier

As the verification article says:

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. (p14)

I checked this concern by doing a statical illustration in the section "The long-term consequence of too much scoring error" focusing on stage 4.5, the Teal altitude. I could do the same here for the MetAware Tier, if I had 1) the actual prevalence of the stages in the population under study, and 2) the magnitude of scoring error for each of the stages. I do not have the actual prevalence, so I cannot do the same calculations. However, for illustration of the important statistical principle here, I can safely say that there will be many more people at 5.0 than at 5.5 in virtually any population. Which means that if the magnitude of scoring error in figure 5 were to be applied on a population of stage 5.0 and 5.5, we would have a very substantial amount of people scored as 5.5 to actually be at 5.0.

This same basic principle applies here too: A high scoring error combined with uneven distribution of the actual prevalence in the population, where there are virtually always much less

people at later stages, will inevitably produce the bias towards an overrating of the number of people at later stages. And the bigger the scoring error, the bigger the magnitude of overrating.

But I will leave the statistics behind and rather look at the two scoring approaches at post-autonomous stages (STAGES MetAware Tier). I will be writing from my own experience as well as from dialogues with a number of people who have an informed experience with EDT and scoring. So far, I have dialogued with at least 7 people who have had the same doubts as to whether one or more persons with the MetAware scoring was really at Turquoise (or later), seen through the filter of EDT. For myself, I had in the past two of my own inventories blindly co-scored by Susanne Cook-Greuter and Terri O'Fallon, and in both instances Susanne gave an earlier center-of-gravity than did Terri.

This means that the concern cited above, in my opinion, now amounts to what is called anecdotal evidence. Which is not to be confused with empirical evidence, which is when we have a proper study, as with the replication study. So my investigation into the scoring system on the MetAware Tier comes from anecdotal evidence and is a thus a qualified attempt to help clarifying where the two approaches appear may differ in terms of the scoring results. The discussion below is not about the distinction in the three latest stages at the MetAware Tier, it focuses basically on the criteria of being scored at the Construct Aware stage (or later). First I will look at where the two scoring methods obviously differ.

The Definitional Differences in the Two Scoring Methods at the MetAware Tier

I have found three differences: How the stages are defined in each model, the fact that STAGES look at whole paragraphs while the MAP does not, and different involved criteria for scoring at these later stages.

First the definitions. In the article we read that

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. … (p. 5)

We see here that STAGES defines some 4.5 Teal inventories to be at 5.5 Unitive, which points to a rather significant difference in how these stages are defined in EDT vs STAGES. This will of course to some extent skewed the STAGES distribution of inventories towards the MetAware Tier, as compared to the MAP.

We also read that 5.0 is "primarily the same as" Construct Aware, and in the very same paragraph that the "two models differ significantly above Strategist". I find this a bit confusing, but the important point here is that it is clear that the definition of Construct Aware differs between the two models.

Secondly, the scoring done by the MAP rarely considers subsequent elaboration in a sentence-stem response. STAGES on the other hand, explicitly states that it does. As the article points out:

> *Sentence completions vary from a few words to full paragraphs, sometimes multiple paragraphs.* (p. 2)

According to O'Fallon (personal communication) people operating at the MetAware tier may need more than one sentence to express themselves fully, and thus the STAGES method looks through the whole response regardless of the number of sentences. According to Cook-Greuter (personal communication) mature individuals are capable of using one single sentence to express themselves in a way that reflects the highest level they have stable and spontaneous access to. The issue here is that the developmental stage next to the one from which we make meaning under normal life circumstances (center of gravity) will to some extent already be emergent. Which means that this next emerging stage will, more or less, be present in the test taker's intrapersonal space. Given more time to reflect deeper increases the likelihood for this next emerging stage to influence the sentence stem response. This means that the STAGES scoring system will, by definition, tend to score some responses at a later stage than the MAP.

In addition to this, I believe that this difference also increases the likelihood of misidentifying cognitive agility with direct metacognitive awareness in STAGES. That is, STAGES gives room for more cognitive elaboration which can introduce more scoring error. More on this in the next section.

The third factor is the involved criteria used in each of the systems. In the MAP, in addition to needing a certain number of responses (the cut point criterion) at these very late stages, Construct Aware has 4 additional criteria (Unitive has 6). In STAGES we have that

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.

I do not know what these stringent scoring rules are for STAGES, but in one of the inventories I had co-scored by O'Fallon and Cook-Greuter the difference in the center-of-gravity they each gave was because of the criteria the MAP is based upon. While I am not saying that the scoring rules used by STAGES are "wrong", the point is that they are obviously different. Again, STAGES differ from the MAP in the scoring rules, and it seems to be that the delineation they have between Strategist and Construct Aware differ in that STAGES is tilted towards scoring more inventories at the Construct Aware level than the MAP does.

Other Potential Differences in the Two Scoring Methods at the MetAware Tier

There are a few other potential differences between the two scoring systems. They can shed more light on the anecdotal evidence I am investigating here.

First, STAGES has a more sophistical model than EDT at the post-autonomous stages, in that they incorporate both states and vantage-points into their theoretical model in how stages unfold.

At the level of the theoretical model, that may be very helpful and useful. As for the scoring system, the question again is how the two scoring systems differ and why they seem to provide different results. The article states that answers at the MetAware Tier can sometimes be expressed as

... experiencing a sense of oneness, life energy, or beauty pervading everything. (p. 4)

As I see it, people could have had a deep state experience either in meditation, or by ingesting an exogenic catalyst, and then talk about a deep and authentic subtle state experience. But that could also be experienced, and articulated, within an ego developmental center of gravity at 4.5 Teal or 4.0 Green. I think the potential conflation between states and stages can happen in both scoring systems, but it appears to me that the MAP is less prone to the conflation. I say this based on my experience of myself and people I know who I believe have been scored at the MetAware Tier prematurely. Many people drawn to these tests are spiritual practitioners with some measure of state training and possible vantage-point shifts. Again, this is more hypothetical, but the methods again do seem to differ.

Secondly, I want to mention the impact of the knowledge of the scoring system itself, prior to doing an inventory. I firmly believe that to whatever extent a test taker is cognitively aware of how the scoring system works, she or he will be more prone to let that influence the responses. And it may not be conscious at all. In the article we read that

...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. (p 5)

A short workshop is thus enough to get an overall understanding of the parameters. The article further explains that the first question the scorer looks at is

Is the response Concrete, Subtle or MetAware? (p. 5)

This means that the STAGES-informed test taker already knows that describing a MetAware object will potentially qualify that sentence to be scored at the MetAware Tier. Again, this influence may very well be totally unconscious and not any deliberate attempt to "game the system". (Gaming the system would mean deliberately writing to get the inventory at a late stage. That can happen too, but that is not what I discuss here.)

I actually caught myself off guard noticing this phenomenon when I was filling out a sentence doing a STAGES inventory. I suddenly realized I was answering from a cognitive knowledge of the STAGES parameters, and not from my direct unfiltered, immediate experience in that moment. I let the sentence stay out of curiosity, took a breath, and kept going. When I got the results that sentence was scored at 5.5. Again for comparison, I have done several MAP inventories over the years, and it is definitely much more difficult to internalize the scoring method for the MAP; that would require extensive study of the method.

I think that the distinction between cognitive knowledge and metacognitive direct knowing is a crucial one, especially at the MetAware Tier. And I have others say the same thing. I'm sure the STAGES scorers know this well too, but the distinction can be very subtle and sometimes hard to detect for anyone. And the better we know any model, the more sophisticated we become at juggling with the concepts involved. Potentially, there could be an increase in MetAware inventories over time because of this.

Again, for our comparison here, I have not seen EDT describe the specifics of how the content in a sentence response may impact the scoring, in the way STAGES do (objects being concrete, subtle of MetAware). It appears to me that the MAP is therefore, relatively speaking, less directly focused on the objects in the sentence and thus, by definition, to some degree more tilted towards the structure in the test takers response. That is of course a hypothesis at this point, since obviously both the MAP and STAGES use both content and structure to score any sentence and I am not familiar with all the subtle details in the two systems.

An Example Indicating Higher Scores with STAGES.

Looking again at the pasted table 3 with the real-life distributions, we see that the two populations to the right (scored by STAGES) have a very different distribution than the first four (scored by the MAP). The STAGES "International self-selected data" from 2014-2018 has 33 % at Teal and 28% at the MetAware Tier. If we compare this with the "International self-selected data" from 2000-2007 based on the MAP, STAGES has 3.7 times as many at Teal altitude, and 8 times as many at the post-autonomous stages (MetAware Tier).

This increase of 370% and 800% respectively, could be explained by development over time, and/or the population the samples come from. However, it is hard to believe these are the only reasons, especially seen in light of the discussion in this article, and the magnitude of the differences. We also see that the increase is higher for the MetAware Tier than for Teal, which is also in accordance with the discussions above. We certainly cannot rule out that the difference in the scoring approaches can have something to do with what we see here. These two quotes from the same book as table 3, the book chapter *The Aware Leader: Supporting Post-Autonomous Leadership Development* (Lynam et al., 2020), are interesting in this respect:

In the early years of the program, we used Cook-Greuter's post-Loevinger research on ego development and the associated Maturity Assessment Profile (SCTi-MAP) sentence completion instrument. In 2014 we began to use the STAGES model and assessment (O'Fallon, 2016), although the STAGES research had been going on for years prior.

A comparison between the data on pre- and post-assessments shows a comparable overall developmental growth in GTC graduates, but in the past nine years, there is more growth in the MetAware stages (5.0 Construct Aware and later). Our interpretation of this data is that GTC is better serving participants that are Strategist and later, as well as attracting more people from the later end of the developmental spectrum.

The increased growth referred to thus started right before the scoring went from the MAP to STAGES.

Final Comments on the MetAware Tier

The discussion in the previous sections came out of an interest to find some explanations for the anecdotal evidence related to the observed difference between the two scoring approaches. But in doing so, I have become increasingly aware of the problems with not properly having a clear distinction between the two theoretical models and respective scoring systems. Remember that the basic claim I started with was that the STAGES objective is to strengthen the ego developmental model and its assessments. As I did for the examination and the subsequent discussion on the replication study, advocating for a clearer conceptual separation of the two constructs, I see the same here on the MetAware Tier, and possibly even more so. As the article points out

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. (p 5)

First, one can arguably say that STAGES has a more sophisticated theoretical model for the dynamics of growth at the post-autonomous territory, what they call the MetAware Tier (especially the three latest stages). More importantly for the discussions in this article, I have shown several quotes which makes it very clear that STAGES defines this territory differently than EDT. Obviously, that makes it virtually impossible to maintain or claim that the two scoring systems would always produce the same results. So we should and must expect differences in the overall long terms stage distributions from the two systems.

One of the main problems I have been facing here, is that STAGES uses the same term "Construct Aware" as do EDT, for the STAGES stage 5.0. But it is not the very same stage the two models talk about or attempts to measure. If we were to apply the Wilberian notion of content-free altitude as a helpful framework here, we can say that "Construct Aware" in EDT (or the numerical labeling 5/6 as EDT also uses for this stage) and 5.0 in STAGES both refers to the Turquoise altitude. In this sense, 5/6 and 5.0 refers to the same altitude, but due to the way the models are constructed differently, they will emphasize different aspects of it as it were.

If this distinction is not made, there remains an implicit belief that "Construct Aware" has an ontological existence outside of EDT. Which it certainly does not, the term was coined by Susanne Cook-Greuter during her work with this and pertains to EDT and the way that model is constructed.

So if someone tells you that "I have been scored as Construct Aware by STAGES", you will (if you have knowledge of this) automatically interpret that through the EDT and MAP lens. And then you may start to wonder about it if it does not fit your experience. If they say "I have been scored as 5.0 by STAGES" that is actually quite different, as 5.0 pertains to STAGES and not to EDT.

In terms of the discussion above on scoring at the MetAware Tier, the question is then how much of the discrepancies between the two models is a result of theoretical differences, and how much is scoring error? Whatever the theoretical difference may be, there is obviously a lot of

scoring error in STAGES, as this article has made a pretty good case for. But in order to make any way forward with the basic conceptual issue, I would need to hear a different narrative from STAGES as to what STAGES is in relation to EDT.

The Positive Responses to STAGES

Many people, including good friends, report having been helped by STAGES. I certainly don't dispute that people feel helped by STAGES. As the article states:

> *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.* (p. 9)

That may be perfectly true, but the issue here is that these anecdotal reports do not say much about the STAGES scoring system as a reliable way to measure ego developmental stages. People also say they are helped by astrology (which is now a 2.2 billion USD industry). My point is that any system putting people into categories will feel helpful for some, regardless of its theoretical underpinnings or accuracy. The question here is the actual ego developmental aspect.

The problem I see relates to a deeper understanding of development as such. My own experience, as well as many people I discuss this with, is that truly grasping the nature of psychological stage-development and its many aspects and varieties requires long experiential training. I consider myself a perpetual learner. We cannot see these stages accurately by direct introspection, and we cannot reflect upon our own stage center of gravity, by definition.

This is in its own way a fascinating topic, but also makes one prone to internalize misunderstandings about the nature of stage-development. When we are told that the STAGES scoring system is measuring ego developmental levels, and even backed up with statistical validation, we will very likely start to internalize this as an objective truth. We will find all kind of reasons why STAGES works (confirmation bias) while we will not be able to detect that it may not work the way it says it does.

I believe this goes for all of us, even if we have studied these models and have the capacity to operate from a late stage ourselves. None of us are too good not to be subject to misunderstandings or internalize wrong beliefs. And the longer we have been subject to it, and created any identification with it, the harder it can be to let it go. If you have been scored by STAGES, the score (whatever it was) may reflect the ego developmental center of gravity, or it may not. What I find helpful is to expose myself to different developmental theories, discuss with peers, stay genuinely curious, lighthearted, and as honest as possible.

With the STAGES scoring system apparently becoming increasingly popular, also due to its inherent pleasant simplicity, I believe it may become harder to detect these types of misunderstandings. A good scoring system will help to expose cognitive biases around development, so that they can be corrected. But that requires sufficient scoring precision which is in accordance with the actual theory itself.

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. The basics of the STAGES scoring system is relatively easy to grasp, and anyone with sufficient interest is potentially able to internalize the basics of the scoring rules to the extent that it will influence the sentence completions and hence the scoring.

Discussion on the STAGES Theoretical Model

I am not in a position to give a coherent analysis of the STAGES theoretical model, as this is also beyond the scope of this article. I am therefore more agnostic about it, and thus open to it being useful and helpful, always in the right circumstance of course. Generally speaking, I am always cautious not to rely too much on a single model, as any model of reality is inherently limited. My experience is that developmental models are often given too much explanatory significance by those who work with them. This situation can be strengthened further since concepts like "development" and "higher stages" are inherently egosyntonic.

In the context of this article, it seems to me that the STAGES theory has been too closely tied to EDT, which then led to the replication study and the wrong validation claim of the current STAGES scoring system. The verification article states that

> *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.* (p. 10)

We could turn this statement around and say that since the STAGES scoring method is empirically validated not to score accurately, it suggests that the validity of the theoretical model as a developmental model is questioned with it. That may have some truth to it, in the sense that the three parameters will have inherent limitations as to what they can alone explain. But it does not say that the STAGES theoretical model is "invalidated". I see the relevant question as about *where* and *how* the STAGES theoretical model is useful. But again that question is beyond the scope of this article.

In the context of this article on the STAGES scoring system, the important conclusion is: Whatever the benefits of the STAGES theoretical model may be, it does not in any way alter the empirical facts of the replication study, or the inter-rater study, on the current STAGES scoring system.

Conclusion

In this article I have presented and discussed the data from the two STAGES studies. The statistics I have applied are relatively basic, and the arguments should be possible to follow for non-statisticians. I want non-statisticians to be able to take part in a dialogue around this issue,

should they choose to. The findings show conclusively that the STAGES scoring system, based on three questions, does not replicate the MAP with sufficient precision.

The root problem with the validation issue is that the label "very strong" has been taken out of context and its meaning misinterpreted. The truth of the replication study is easy to see when we take the quality standards for scoring precision into account and look at figure 2 directly. It is not statistically valid to churn all this data into one single number and then look into a statistical textbook to find a validity criterion there.

We have also seen that the replication study reveals that the STAGES scorers do not reliably match the master scorer's own STAGES scores (figure 3). This is a different but important result, and it stems from the same underlying magnitude of scoring error we can observe across the measures of the STAGES scoring methodology. I believe, as I have outlined, that the scoring error is due to the fact that the entirety of the MAP protocol simply cannot be reduced to the three questions. There is some explanatory power in the three questions, but they do not replicate the MAP. Simply stated: The ego development construct cannot be reduced to three underlying parameters.

The numerical results are not my personal opinion, they are what the studies themselves reveal. I have only used the data from STAGES' own studies, the information in the validation article itself (and a table from a book chapter produced by STAGES proponents) to put the pieces together.

I hope these results have been made clear by this article, and that there can be an agreement about what the studies reveal. I hope that what I have shown will be recognized by the STAGES community and acted upon without unnecessary delay in a compassionate, skillful way.

I would also recommend a careful examination of other studies done on the STAGES scoring system or the STAGES theoretical model, especially to the extent they are based on, explicitly or implicitly, the assumption about ego developmental stage replication as we have in the STAGES verification article. I have looked into some other STAGES studies briefly, but wherever they are based on data from STAGES inventories, of course I cannot say much about their overall validity.

As a purchaser of a product, in this case an assessment that portrays to measure my ego developmental center of gravity, I cannot trust the current STAGES scoring system. From that perspective, STAGES must come to terms with its unsupported narrative to be an extension of ego developmental theory and an improvement over the MAP assessment system.

In my opinion, one needs to adequately distinguish the STAGES scoring system from the underlying STAGES theory. If not, and the current STAGES scoring system is not corrected, the chance is that the STAGES scoring system will undermine the STAGES theory along with its potential benefits.

Finally, based on my own investigation into this, I see the root problem being that STAGES has tied itself too closely to another developmental construct. The STAGES theory does not retain the base of EDT, and the current STAGES scoring system does not replicate, let alone strengthen,

the MAP. If STAGES comes to terms with the actual results from the replication and inter-rater studies, and properly differentiates itself from EDT, I believe it will reduce a lot of confusion, as well as tension, around this issue. And I firmly believe it will only help STAGES in the long run.

Appendix

Previous STAGES Critique and Rebuttal

The earlier critique of STAGES was published in 2017 in Integral Leadership Review. The interested reader can go here. For the rebuttal from the STAGES community, in the same ILR issue, go here.

What is interesting to be aware of, especially seen in light of this present article on the replication study, is the way the rebuttal answers the need for rigor by referring to the replication study. Note that this was before the study and analysis were published. Here are a few quotes from the rebuttal:

> *The replicability of all four STAGES scorers…. was found to be in the "excellent" range. Thus confirming the primary hypothesis of the study.* (p 10)

> *… this summary is not meant to be a sufficient argument for the validity of the STAGES model or scoring system. This summary merely points to a to-be-published paper containing sufficient detail to back up such a claim.* (p 10)

> *We have granted that there was a period during which O'Fallon and colleagues made public claims too far before publication of results. However, the research design and analysis have been rigorous.* (p 10)

The Role of the Statisticians

I will also quote from the rebuttal some words from Nayak Polissar, the lead statisticians behind the replication study:

> *I have worked on many studies comparing methods for scoring or classifying people or phenomena, and the agreement in this study (based on a statistically substantial number of inventories) is very impressive…. I am not conversant with developmental theory, so I do not know how to evaluate whether or not Dr. O'Fallon's theoretical justification for her method fits well with existing theory or with methods for developing new theory. I do know that she has something that works. The study results will show that STAGES scoring does well in matching up with the Loevinger/Cook-Greuter scoring method. …. The method works.* (p 14)

The issue I see here is that Polissar does not seem to be in a position to claim that "the method works" the way he does, since he reveals a limited understanding of the theory that the method is based upon. This is pretty clear when he says that he is not "conversant with developmental theory." Nayak and I have had some good and respectful conversations, and I fully trust his and his team's capacity to do all the mathematical calculations very well, so I am not questioning that aspect of this at all.

One way to shed more light on this is to realize that statistics is always a support discipline. It is therefore important to have access to the basic principles of the field you are supporting. Without this, you can easily end up working with statistical construct without direct relevance to the actual problem at hand. I call this phenomenon "lost in statistics".

I want to emphasize that I do not attempt to make Polissar into the scapegoat here. The statistician is very dependent upon the accuracy of the information he gets from those who hire and work with him, in this case Terri O'Fallon and Tom Murray. I do not know anything about their dialogues during this process.

I strongly maintain that in an analysis of a replication study, it is important to factor in the inherent quality standard in the scoring methodology being investigated, which I have explained at length in this article. And that has nothing to do with the mathematical aspects of statistics. It is all about where you focus your attention based on your knowledge of the particular field you are investigating.

My professional approach in this particular case goes like this: First you check if the basic inherent quality standard in the method is met or not. If yes, you can go on to a variety of deeper analysis depending on what you want to investigate. If not, you simply stop and correct the basic inaccuracies in the method before you do anything else. It is similar to Occam's razor: The simple display in figure 2 and 3, coupled with the information about the quality standard in the scoring method, is sufficient to understand that the verification claim in the article is invalid. (If you need numerical proof of this, you can go to the last section in this appendix.)

To unpack the statistical claims behind this article has been an important motivation for me during this investigation. It is virtually impossible for non-statisticians to detect the statistical fallacies in the arguments we see in the STAGES verification article. I therefore hope I have achieved my aim of making this visible to the non-statistical reader.

In the end, the question in case here is not which statistician you should listen to only based on our mathematical capacities, earlier work, or whatever other preferences you may have. The question here is: Do you think that the 85% direct match criterion, explicitly stated by STAGES, and not more than one stage deviations, should be factored in, or do you disagree? This is up to the STAGES community to decide upon, and the final word should not be left to statistical professionals. That would be to forfeit this important decision to the statisticians, and for STAGES to abdicate their authority over their own method. STAGES makes a choice, and then states that choice publicly.

The Scoring System

The final stage score that both scoring systems produce is the result of individual scores from 36 sentences. Each of the 36 sentences is given a stage score, and the final stage score for the whole inventory depends on all these 36 individual stage scores.

Note here that the 36 sentences form a *distribution of stages* in the inventory, from which the final stage is deduced. In ego developmental theory, this final stage, also called the center of

gravity, is the stage from which we make meaning under normal life circumstances. The earlier stages do not go away, which will be seen in the distribution of the stages from the 36 sentence completions. Further, later stages may be emergent and present in the inventory, but they have not yet been metabolized to the extent that the general meaning making has shifted in the individual. This to say that we need very high precision for each of 36 sentences to deduce the right final score. And right score means in accordance with ego developmental theory itself.

The cut off rules for determining the final stage score by STAGES is described in the STAGES article, but I will not go into those specifics here. I will rather explain what we see, generally speaking, regarding scoring accuracy on individual sentences.

So, in order for the final score to be accurate and in accordance with the given theory, we must have a high and reliable precision for each of the 36 individual scores. When we have that, we will over time meet the minimum 85% direct match criterion for all inventories, and very rarely, if ever, miss by more than 1 stage for any inventory. For the MAP this is achieved by relying on the manual of sentence completions (which implicitly contains within it ego developmental theory), as well as the addition of 14 different criteria.

If the probability increases for the individual sentences to be scored incorrectly, we are introducing more scoring error into the final inventory score. In order to have a deviation of more than 1 stage on the inventory, there usually needs to be several wrong scores among the 36 sentences. The replication study has revealed that the STAGES scoring system has too many wrong scores on whole inventories, including many deviations of more than 1 stage. This cannot be anything but the effect of the lack of sufficient precision for scoring each individual sentence correctly, that is, in accordance with the underlying theory itself.

So, the lack of precision in the STAGES scoring system, as revealed by the replication study, is a natural consequence of too much scoring error on individual sentences. That lack of precision proliferates through the inventories and the current STAGES scoring system breaks down as a reliable instrument.

Combining the Estimated Results from Figure 2 with Real-life Populations

The estimated probability of scoring someone who is *at* 3.5 (by the MAP) *as* 4.5 (by STAGES), will be 3 out of 24, which is 12.5%. We see this in figure 2, where the number in the cell with 4.5 horizontally (STAGES) and 3.5 vertically (MAP) is 3, and the total number under the 3.5 column is 24. That is, out of 24 inventories at 3.5 by the MAP, 3 is scored as 4.5 by STAGES. For 4.0 we have 7 out of 24 which is 29.2%, and for 4.5 (correct score) we have 14 out of 24 which is 58.3 %. (I will not add the 2 very strange STAGES scores as 4.5 which was scored at 2.5 by the MAP).[4]

[4] Note that these numbers are estimated probabilities with inherent statistical random noise, but they are unbiased estimates coming from a solid study, so they are real in this sense. Since we do not have estimated probabilities of scoring as 4.5 by STAGES for EDT stages later than 4.5, we must constrain ourselves to the stages 3.5, 4.0 and 4.5. It would hardly alter the results since there is much less prevalence of these later stages in any population. And it would only make the probability of correct mapping less.

The real-life populations from table 3 I have copied into figure 4a, below. (For population 4 we only have 72% for all stages earlier than 4.0, so I chose 30% for 3.5. It is lower than for all the other 3 populations.) For the probabilities I use standard statistical notations for conditioned probabilities.

		Real life populations			
Probabilities		Pop 1	Pop 2	Pop 3	Pop 4
P(4.5|3.5)	12.5%	33.5%	34.8%	36.5%	30.0%
P(4.5|4.0)	29.2%	23.4%	5.0%	11.3%	15.5%
P(4.5|4.5)	58.3%	13.5%	1.4%	4.9%	9.0%

Figure 4a. Real life populations.

The next step is to multiply the probabilities with the actual population size for each stage from each population, which gives us figure 4 b. For instance, for population 2 we have 12.5% * 34.8% = 4.4%, which is the estimated percentage of people in that population *at* 3.5 *and* scored *as* 4.5 by STAGES.

	Percentages scored as 4.5 by STAGES			
MAP stage	Pop 1	Pop 2	Pop 3	Pop 4
3.5	4.2%	4.4%	4.6%	3.8%
4.0	6.8%	1.5%	3.3%	4.5%
4.5	7.9%	0.8%	2.9%	5.2%
Total:	18.9%	6.6%	10.7%	13.5%

Figure 4 b. Percentages being scored as 4.5 by STAGES.

We see in the bottom row Total in figure 4 b that the percentages of people scored as 4.5 by STAGES is around 5% higher than the actual percentage in each of the 4 populations (which we find in the bottom row in figure 4 a). For example, the percentage at 4.5 (Teal altitude) in population 3 is 4.9% figure 4 a), but those scored with STAGES *as* 4.5 from this population will be 10.7%. We see that the anecdotal concerns noted above are pointing to something real.

We can finally scale up to 100% for each population to make it easier to interpret what we are looking at. We now get the distribution over the actual stages 3.5, 4.0 and 4.5 for those scored *as* 4.5 by the STAGES assessment. This gives us figure 4.

The Difference Between 2.0 and 2.5

These two stages are respectively called D3 (Delta) and 3 (Diplomat) by EDT. The article says the following about the stage 2.0 (D3):

> *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). (p8)

But the reason stage D3 is subsumed in stage 3 is solely for practical purposes, due to the rarity of D3 in most adult populations. In the Cook-Greuter document referred to in the introduction in this present article, we read about D3 and the MAP that:

It does not eliminate this level when actually dealing with populations in which it is of theoretical importance. (p 122).

The STAGES validation article further states that

...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. (p6)

So the MAP stage 3 inventories are mapped to 2.5 because of conceptual similarity, but some could potentially be 2.0 (D3). And we know that the numbers of 2.0 out of these will likely be very few, due to the rarity of this 2.0 stage. I have expanded figure 2 by splitting 2.0 and 2.5, which gives figure 6 below.

		MAP Scoring									
		1.0 (2)	1.5 (2/3)	2.0 (D3)	2.5 (3)	3.0 (3/4)	3.5 (4)	4.0 (4/5)	4.5 (5)	Total	Agreement
3 INDEPENDENT SCORERS	1.0	**2**								2	100.0%
	1.5	4	**2**	2		2				10	20.0%
	2.0	2	3	**6**	15	8				34	8.8%
	2.5		1		**1**	1	1			4	25.0%
	3.0		2	2	6	**10**	1			21	47.6%
	3.5					3	**17**	6	2	28	60.7%
	4.0						2	**10**	4	16	62.5%
	4.5				2		3	7	**14**	26	53.8%
	5.0							1	3	4	
	5.5								1	1	
	Total	8	8	10	24	24	24	24	24	146	
	Agreement	25.0%	25.0%	60.0%	4.2%	41.7%	70.8%	41.7%	58.3%		**42.5%**

Figure 6. The 3 STAGES scores with 2.0 (D3) and 2.5 (3) split.

Here we see that out of the 24 scorings by MAP at 2.5, 15 ends up (by STAGES) as 2.0, while *only 1* ends up as 2.5. Some of those 15 may have been 2.0 (D3) subsumed into 2.5 (3) by the MAP, but due to its rarity most likely only a few. What one sees is a very strong tendency of STAGES to score stage 3 MAP inventories not as 2.5, but as 2.0.

Keeping 2.0 and 2.5 separated as they are presented in the data we have available, we go from 52.7% agreement when 2.0 and 2.5 are merged, to an overall agreement of 42.5 % when they are split. And the magnitude of 2 stage (or more) mismatches goes from 11.0% to 18.5%. Assuming a few of the 2.5 inventories from MAP were actually subsumed 2.0's, the percentage of overall

agreement will be slightly higher, but probably not more than 45%. The magnitude of 2+ stage mismatches, however, remains 18.5% regardless.

In other words, the full representation of the replication study, performed by certified STAGES scorers not informed by the MAP, gives an overall success rate of less than 50%, and 18.5% chance of 2+ stages mismatch. A wrong score is more likely than a right score, and more than 1 out of 6 inventories misses the mark with 2 or more stages.

The Probability of Getting Figure 2 by Chance

A simple way to test the replication hypothesis statistically is to treat the scores as independent trials with a constant success probability of 0.85. This is a very good approximation, since the sample technique is so well done. Any possible minor deviations will not have any impact on the conclusions. To be on the absolute safe side I have in addition postulated the maximum random noise in the original MAP scores by treating them as having 85% chance of being correct. Assuming independence between the two scoring systems is what gives the *minimum* probability of 0.85 * 0.85 = 0.7225 of a correct match.

We can now calculate the probability of arriving at figure 2, *if the 85% criterion is met*. Using the central limit theorem to approximate the binomial distribution, the command in R (the same program the statisticians used to provide the numbers in the STAGES validation article) is:

pnorm ((77-146*0.7225)/sqrt(146*0.7225*(1-0.7225))) with the result 0.00000000701

But it may statistically be more correct to use the 85% match without correction. The command is then:

pnorm((77-146*0.85)/sqrt(146*0.85*(1-0.85))) and the result is
0.000000000000000000000000000048

Now, *that's* a small number and a good way to end this article.

References

Lynam, A., Fitch, G., & O'Fallon, T. (2020). The aware leader: Supporting post-autonomous leadership development. In *Maturing leadership: How adult development impacts leadership*. Emerald Publishing Limited.

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

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