Final Comments on the STAGES Validation Claim

Kristian Merckoll[1]

Introduction

I want to thank Murray for his response and bringing more clarity about his perspectives. In fact, his response answers many of my questions about the STAGES scoring system and I will speak to those below. But first I want to remind the reader about why I am engaged in this dialogue. My professional feedback was actively solicited by O'Fallon (in 2019, after the ILR rebuttal); I was shown a version of the verification article before it was published, asked to give feedback and I gave some general comments. Worth mentioning is that I recommended taking the references to AQAL out, examine the discrepancies between the two scoring systems (table 5) and pay attention to the fact that the statisticians were not conversant with developmental theory. After all my subsequent conversations with the authors of the article we arrived at an impasse, as I did not experience that my concerns were properly addressed, except the one about AQAL.

This then made me choose to present my arguments to a larger audience. What I wanted was to unmask the statistical claims and to make it possible to understand for non-statisticians, which I can do since I have 30 years of professional experience. Therefore, the language is not technically or statistically sophisticated, for instance when I use terms like "quality standard." I did not attempt to go into or elaborate on important terms like validation, internal validity, external validity, inter-rater reliability and so on. I do admit however, that I could have made this more explicit to avoid confusion around terms like "validity," which we use colloquially but has a specific meaning in statistics. Murray is using a lot of this statistical jargon in order to rebut and invalidate my arguments, but what is essential for me is to shed light on the specific *claims* made by STAGES, and the way they are using empirical data and statistical methods to verify them.

Before proceeding, I also want to reiterate strongly that I am not against the STAGES theory. Nor I am I against scoring systems; I said in my article I say that "*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*."

In the ILR rebuttal I referred to in my article, STAGES writes the following:

> *Both as defenders of STAGES and as owners of businesses, we do not claim, implicitly or explicitly, to have an objective viewpoint on these matters. This undoubtedly introduces*

[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

biases that we invite the larger community to comment upon…. We ask the community to "keep us honest" in this regard with frank feedback.

Since STAGES is putting forward a new scoring system, telling the public they have empirical evidence for it, then soliciting professional feedback from me, I am doing my best to give that feedback. STAGES is honest about their own biases and invites all of us to give honest feedback. I find this a very reassuring move from STAGES.

In this final piece from me I will not go into all the arguments from Murray which are not directly relevant to my own article. They may have relevance for further discussions, however, so I am not dismissing them as such.

The Distinction Between a Theory, its Application and a Scoring System

In order to continue to discuss the STAGES scoring system, and the empirical studies about it, it is crucially important to keep the distinction between the theory and the scoring system. I do that in my article, but I want to repeat it here. So again, please know that I am not against the STAGES theory, nor its applications. I see it as a very interesting theory with some measure of real explanatory power, especially in conjunction with other theories. I have used it myself in collaboration with people who use it skilfully. But that does not necessarily *require* a scoring system. An adequate scoring system can help, but if the scoring system has too many frailties, it can end up undermining our understanding and skilful application of the theory.

I will continue to focus on the STAGES scoring system, and only refer to the theoretical model where necessary for clarity.

The STAGES Scoring System Does Not Attempt to Give the Same Results as the MAP

As Murrays says:

To clear one thing up at the start, we never claimed that STAGES scoring will give one a facsimile of a MAP score, or that it is a new or better method for getting your MAP score. Though in hindsight this fact could and should have been emphasized more in the Heliyon paper. …. we have never said that "if you receive a score on a STAGES assessment it should match well with what you would get on a MAP assessment." …So as to Merckoll's claim that "the STAGES scoring system is not fully validated as an alternative to the MAP," we agree that it is not, nor was it ever intended to be, an alternative way to score a MAP inventory.

As I far as I am concerned, this is a major shift in how STAGES relates to their scoring system. It is quite different from the impression I have from the verification article, and from all the people I have discussed this with over the years, including Murray and O'Fallon. I am very appreciative of this clarification from Murray, as this fact is now explicitly and clearly stated. One important implication here is that whoever has gotten a STAGES inventory, are using it professionally, or plan to buy one in the future, are now confident in knowing that it may not yield the same stage

result, centre-of-gravity, as the original ego development scoring system would have given. Another implication of this is that since the MAP (or any other version of the same sentence completion test with manual of sentence completions and all the other criteria they use) reflects the developmental stages in EDT, the STAGES scoring system does not attempt to measure those. Hence, STAGES has now properly differentiated itself from EDT (as an ego developmental model) as I suggested in my article: "*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."*

So, we can now put this central point behind us, and from now on, I will approach STAGES as a related, yet different, developmental model. There remains however, some general credibility issues for me, since the validation article says that "*STAGES successfully replicates the prior Cook-Greuter/Loevinger scoring system to a "very strong" degree up to level 4.5*." I hope that STAGES realises the impact this language has had, and some of the confusion it has created.

The Training the 3 STAGES Scorers Received

Murray states that the 3 scorers at the time of the replication study received a training that *was extremely minimal compared to what later emerged. …. the scorers who participated in the Heliyon study had nothing even closely approximating this level of training (they had less than one day of introduction to the methodology, based on preliminary scoring rules); and the scoring manual was not yet created.* They were "*green,*" the training was "*ad-hoc*" and there was no "*certification program*" at the time.

I do not get any impression of this reading the validation article; it states that the 3 "*were the first certified scorers of the new STAGES system.*" Murray says that "*Merckoll's point that errors in more frequently seen levels have greater potential impact is reasonable, but the inter-rater match from the Heliyon results do not generalize to overall trends because of the minimal scorer training we have mentioned."*

Me giving professional feedback on the empirical validation has to relate to the data at hand. I cannot dismiss that and start speculating about what may come later. That is not being biased against STAGES and I do not think Murray's labelling of my explorations as "moot," "specious" and so on are fair. I have simply been looking at the data presented in combination with the specifics of the validation claim and the other information presented in the verification article. I wanted to make it possible for non-statisticians to understand that this claim is not warranted, given the evidence presented. I took the time to make additional calculations, and to discuss possible explanations as well as consequences.

Murray says that: *Because Merckoll extrapolates erroneously and at such length with speculations about the STAGES scoring system's quality as it exists beyond the Heliyon study, we need to set the record straight.* But I do not speculate beyond the Helion study, I only look at the empirical evidence at hand and make it crystal clear what *would be* the impact with this level of discrepancies between the two scoring systems. The validation article makes many very strong assumptions about the benefits of the STAGES scoring system and uses those very same data to justify it. According to Murray, I should not even be allowed to examine his own published data

critically, since the 3 scorers were so poorly trained. I find this attempt by Murray to discredit my solicited feedback disingenuous. What I have achieved though, is to have Murray clearly state was actually behind the replication data. I accept that from Murray, but not his labelling of my arguments and explorations. But since Murray never meant to say that a STAGES assessment should match well with what you would get on a MAP assessment, we have already moved past one of our main disagreements.

The question remains if the new training is providing better inter-rater reliability. I find that likely, since the results from the replication study were so poor. But we can never know for sure, as the ILR rebuttal (2017) says (quoted in my article): *it is always possible that future research will not show results as strong as the recent replicability study.* Given new real evidence, I am of course totally open to let that inform me. Murray says that they *have also published other papers on the inter-rater reliability of STAGES scoring*. I have looked at them and will comment below.

The Application of Statistics

In the verification article, the claim is tentative support for the STAGES theory, through the replication study with MAP results. If we go some years back in time, before results from the replication study were published, STAGES said that the study was a statistical proof of AQAL.

But we do not need any statistical proof of the STAGES theory, let alone AQAL. It doesn't work that way, and I see our disagreements here due to a very different approach towards verification, and what the domain of statistics is in this case.

I'll try to explain it this way: Cook-Greuter came up with the increasing person perspectives as a way to understand the unfoldment of the stages in EDT. For example, a 3rd person perspective comes before a 4th person perspective, and for each of them early comes before late. The sequence is early 3rd person perspective, late 3rd person perspective, early 4th person perspective and late 4th person perspective. In STAGES, these four constitute the subtle Tier and are labelled this way: passive singular, active singular, passive (or reciprocal) plural and active (or interpenetrative) plural. I don't need any *statistical proof* to see the value and usefulness of any of those distinctions. What I need is a basic understanding of how they are defined and what they mean.

In Murray's response he writes about the stages in EDT that: *There was, prior to STAGES, no concise way to describe each or any of them, never mind the trajectory through the entire sequence.* For me the increasing person perspectives from Cook-Greuter is a very useful, concise, and helpful description.

As for AQAL, it goes like this: When I say "I", I actually mean me sitting here. When I use the term "we" I am referring to me and somebody else, in this case us engaging in this inquiry together. When I use "it" language in a 3rd person fashion, I am referring to the objects we discuss here, like the STAGES scoring system. Do you need me to prove this to you statistically? Or do you want me to prove to you statistically that a cell contains molecules, but that a molecule does not contain cells? I better continue to stick to the scoring system.

What is STAGES Measuring?

In Murray's response we read that: "*The STAGES system is not another "rater" for the MAP system, it is a completely different technique for assessing what we hypothesized to be approximately the same psychological construct."* And that:

> *The STAGES scoring system offers a radical proposal: that most of the complexity described in ego development methods is the result of just three underlying variables or parameters.* And further: *Our research question is whether the two systems seem to be measuring close to the same underlying psychological construct, by comparing the final "center of gravity" scores (aggregated over 36 completions) from each system.*

What the STAGES scoring system is attempting to measure is the three underlying variables and should be treated as such. As for the proposed *same underlying psychological construct*, I relate to EDT and the concurrent sentence completion test this way: The test and the theory have in large part co-evolved, and we need not define one single underlying psychological construct in order for it to work and have value. The value and usefulness are in the many descriptions associated with the different stages, which have been accumulated over more than 50 years. There may be underlying psychological constructs which can be defined, but all the useful details in the descriptions associated with each stage cannot be reduced to them. In the same way as the properties of water cannot be reduced to the properties of hydrogen and oxygen.

Does "passive" pertain to an underlying psychological construct, while "early" is a surface description? I'm sure my notion can be challenged, and I am not categorically dismissing the radical proposal, so I'll leave this discussion open. I align more with Polissar here, who in the ILR rebuttal said that he does *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.*

But whatever can be said to be the real case about an underlying psychological construct, the explanatory power of using the STAGES parameters to build an unfolding sequence of stages stays the same regardless. EDT and the MAP stays the same regardless. For my analysis, nothing has changed.

How to Approach the Data with Statistics

I have been pondering over Murray's extensive elaboration on the statistical side of things. I have reflected back on my conversations with Polissar and reached out to him again a few weeks ago, but I haven't had any response. So, I have been wondering how to approach this from my point of view, being objective and respectful, but avoiding falling into a non-sensical battle between statisticians which no outsider can follow. This is what I wanted to avoid with my article, but now Murray is bringing that back in full force, using his own statistical understanding. Murray tells us that *our statistician is an expert at psychological (psychometric) analysis, and there is nothing special about developmental psychometrics here (as opposed to, say, personality or instructional design assessment).*

On my side however, I am going underneath the psychometrical and statistical arguments to see what I can find. Knowing the original MAP assessment well, having studied it, used it, and worked with it, I let that inform me as best I can. I have also studied STAGES to some extent, I even have a diploma from O'Fallon somewhere. As for the statistical approach towards verification in this case, I apply my understanding of how the sentence completion test (which the MAP is based on) is designed to work. What does it measure, what does it not measure, what standard have they set for inter-rater reliability, what is the practical impact of more than one stage deviation in an assessment, and so on, as there are many details to pay attention to. It has become very clear to me that a high degree of reliability is crucial, and that missing the assessment with more than one stage must be avoided. These quality standards are intrinsic to this particular assessment and reflects an important property of it. From there, depending on what the question is, and what the claims are, I look for the appropriate statistical tools, and I attempt to make it as simple as possible. (As Einstein said, as simple as possible but not simpler).

To sum it all up again: The high percentage of stage discrepancies we see in the replication data, displayed in my figures, and especially all the more than one stage differences, tells us right there that something fundamental is not in place. The sentence completion test just doesn't work that way, and claims that STAGES strengthen the ego development assessment has no basis in these data. When I realized that, I had no need to make any further statistical analysis, I only wanted to make it explicit for all to see.

As for overrating over more than one stage, I want to warn against it again, as I did in my article. When we are in an individual coaching context, as one of my developmental coach friends told me, a miss of two levels would make the coaching endeavor nearly impossible. Overrating is a much greater disservice to a developmentally oriented coachee than underrating. And that is because an overrated person may view the work they actually need to do as beneath them, or too simplistic.

When I asked Polissar about factoring in the criteria for reliability his answer was: *The bottom line for me is that when we hear a report of 85% accuracy or any level of accuracy, I don't know what it means.* And that it *is of questioning value without further information* (e-mail, Jan 5th 2021). Murray is aware of these conversations, yet in his response now, Murray says that *Our statistician was well aware of the 85% match criteria for inter-raters and decided not to factor that into the analysis as it was not necessary.*

My understanding here is that since the professionals behind the verification article are not conversant with developmental theory, they will approach it as another psychometrical tool and revert to the instruments they are used to applying. Murray says that his statistician *is a senior practitioner in the field*. But which "field" are we talking about then since it cannot be developmental theory? The field I am concerned about is the developmental construct called ego development (one version of it) and its measurement, and I do not need to go outside of that field. I am not against using psychometrics, but the way it has been applied here has led to unrealistic claims, and it is simple to point that out. Those claims, uncontested, create confusion and problems down the road. It will not help the field of developmental theory going forward. But once the claims are reworded, as Murray now has done about the relationship between STAGES and ETD/MAP, we are over in a new discussion which is much more fruitful. And I think it is much better for everybody.

My understanding of the feedback I got from Polissar is that he sees my approach as too unscientific. Be that as it may, he takes the data from the replication study, puts on statistical makeup, wraps them up in a nice-looking Kappa dress so that it looks scientific, and tells all of us that the STAGES method works. As he said to me, his own research approach on the STAGES scoring system "*is usually sufficient license for a new method to go out into the world*" (e-mail, Jan 5th, 2021).

Validity and Convergent Validity

Statistically speaking, validity refers to the accuracy of a measure, while reliability to its consistency. But with the new information from Murray, we are actually not looking at adequately trained scorers, and we are not looking for direct matches due to the difference in methods. So then what are we actually doing here and how should we properly relate to the data from the replication study? This explanation from Murray is illuminating: *There was no need to develop a manual and certification program (both quite sizable tasks) until and unless the basic method was proven to be valid empirically – which was the purpose of the replication study.* So, is the basic STAGES method proven to be valid empirically, just with data from scorers with very little training?

In my figure 6 we see that the clear tendency is for STAGES to score the stages 1.0 and 1.5 higher than the MAP, and 2.5 and 3.0 lower and with many instances of more than 1 stage difference. Again, since there are poorly trained scorers and not the same method, it's a little tricky to say much here, since there is more than one source of rating discrepancies. But it did not look any good for validity since any decent scoring system needs to capture each stage adequately, and not just on average over the stages. I am not sure that if you have the head in the oven and the legs in the freezer, you are doing well on average. But as for validity in the sense that STAGES would score the same as the MAP, this is no longer the question anyway.

About how to relate to the differences between the two systems given our data, Murray says that "*there is no way to tell how much of the variation is due to explainable differences in the models/methods, vs. how much is due to scorer "error" or subjectivity."* But then what about my figure 3? That figure measures the 3 independent scorers against the master scorer O'Fallon, and they *all use the same method*. I made that figure exactly to answer Murray's question, and I explained it in my article. But then again, because O'Fallon was trained using the MAP, and Murray agrees that *some degree of this prior influence is possible and even likely* we have yet another problem with interpreting the data accurately.

Murray says that: "*In fact, it is notable that Merckoll did not find any systematic pattern (such as a consistent offset) in the differences between the new scorers and master scorer, but rather that the differences seemed "random.""* I was not looking for that, and why should I? If I did say that it seems like stage 3.0 tends to get scored as 2.0, Murray could just say that is due to poor scorer training. Murray can look at figure 3 and see for himself. What I did find however, also in figure 3 and not just in figure 2, was way too much variety than could be explained by chance, given the STAGES own stated reliability measure. Remember that the headline of the validation article is: "*The validation of a new scoring method for assessing ego development based on three dimensions of language*." And in the validation article "*This study validates the STAGES scoring method."* But

in order to have overall *validation*, we need both *validity and reliability*. Without reliability there is no validation.

I cannot say that the STAGES scoring system today does not have validity in the sense that it does not measure the three parameters more or less accurately (even though we need reliability too). What I know now however, is that it does not necessarily measure the same as the MAP, hence it does not attempt to measure the stages in EDT. Therefore, that part of my exploration on the validation article is done. As for using the Kappa as Murray does for validity here, see the next section. And for further studies which will eventually shed more light on this, I discuss that below too.

The 85% Criteria

In Murray's response I read that:

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

First, I never said that those two tasks (reliability and validity) need the same criteria. I said, in my colloquial language, that if you do not have reliability, you have no overall validation. But probably most importantly is the magnitude of more than one stage difference, than the 85% criteria on its own. We have seen that the replication study reveals an exact match of less than 50%, and that there are 18,5% mismatches over more than 1 stage. Is that a "very strong" agreement? What Murray is actually saying is that the "very strong" label from the Kappa statistic takes precedence, and that he is therefore satisfied with all these deviations.

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

The replication study is now reduced to an external validity study, and the stringent exact match criteria, coming not from me, but from STAGES, was not really the issue anyway. Again, look at the claims put forth by STAGES in the validation article. Murray further says that:

> Though both methods explicitly commit to the 85% exact match standard for scorer certification, we know of no published or publicly available study or data that verifies this for the MAP system and its trained scorers.

I will share something that may help shed some light on this. This is not any available published empirical study, but I got permission to share the summary data I refer to below. After I submitted my first draft of my article (May 2023) to Jonathan Reams (the editor here) he put me in touch with a friend of his who was looking for statistical help to interpret inter-rater data. These were not from VeDa (MAP) but based on the L/C-G lineage methodology with exemplar based manual and the same set of several additional criteria. They had so far 44 blindly doubled scored inventories, randomly chosen from their ongoing scoring, including 5 scorers. I found an exact match of 86,4%, and not one single instance of a deviation of more than one stage. This is exactly as expected and nothing to write home about. Why? Because it is simply how this particular measurement system works. The 85% exact match criteria and no deviations over more than one stage simply *reflects* the reliability of this particular measurement. Once this quantitative replication box was ticked, one can go on to more in-depth analysis, in particular the qualitative TPR criteria which is an important part of these assessments.

Relativizing Terminology

When Murray relativizes the terms we discuss, we are facing a problem. In the rebuttal in the ILR issue, we read that:

> From an integral perspective we can move away from simplistic notions of "objectivity" and "proof" and address the practical nuances and tradeoffs, and see these tensions in terms of ongoing processes and "polarities to be managed" rather than "problems to be solved."

I find this a very problematic stance. Invoking an "integral perspective" does not make the notion of "objectivity" or "proof" simplistic. This is not "integral," which is all about transcending and including, not transcending and excluding. I will warn against this tendency when we are discussing empirical evidence based on statistics.

In his response Murray writes that:

> No one has claim on the definition of ego, and not Loevinger, nor any other scholar, can or has given a definitive and precise definition of ego or ego development. But these concepts are important enough to our understanding of the human condition that we try to define and measure them.

As far as I am concerned, the L/C-G lineage is in fact giving their definition of the term ego, so that it can be worked with. It is not the only valid definition, but it is clear enough for their purpose.

Is STAGES measuring another variety of what can be labelled ego developmental stages? I leave that to others to discuss. What is important here is that those STAGES' stages are not exactly the same as the EDT stages, according to Murray himself.

Application and Interpretation of the Kappa

Murray writes that:

> *Put simply, if you use the Kappa statistic the results are great, if you use the exact match statistic suggested by Merckoll, the results do not look good! Now our task is to explain why the Kappa statistic is the appropriate choice in a way that the non-technical reader can appreciate." As noted, using the Kappa statistic commonly used in such situations, the match was "very strong." As noted by Merckoll, if one uses an exact matching criteria that gives close misses no credit, the alignment is much worse. So which method is right?" ".. we did not do the extra work to notice that the IRR of our Heliyon scorers was that much worse than our own standard for scorer certification." "the inter-rater reliability based on the weighted Kappa statistic was quite good."*

I find some real problems with this line of reasoning. First, the results do not look great because of a high Kappa, it only looks great because of the label attached to it. I'll say more about that below. And please, as I have said to Murray before, I do not suggest an exact match "statistic," and it is not a "method." No method is "right" here, it has to do with how we interpret the results we get. The 85% criterion is not my idea and has nothing to do with statistics. It is what STAGES says about their own standard for reliability! And you cannot let a Kappa or any other measure override that; if you do it's a serious misuse of statistics. Murray is free to set that standard as he wants, but he cannot maintain an 85% criterion and keep on claiming "very strong" results as he does here.

> *The Kappa statistic is used in many types of comparison studies, is often used for inter-rater studies, and is sometimes used for convergent validity studies. It is considered a more rigorous method than alternatives because it takes into consideration that some scores might match simply from random chance. We can provide evidence from the literature that the kappa statistic is appropriate for convergent validity, while Merckoll gives no references supporting his opinion that it should not be so used.*

Since Murray refers to the Kappa quite a lot, I will explain my take on it with some references before I close the chapter on Kappa. I will discuss the impact of the a priori relationship between the two scoring systems, considering the necessary number of exact matches, and the difference between replication and validity.

From the article *The Kappa Statistic in Reliability Studies: Use, Interpretation, and Sample Size Requirements*[2]: "It indicates the proportion of agreement beyond what is expected by chance, that is, the *achieved* beyond-change agreement as a proportion of the *possible* beyond-change agreement." What this means is that any a priory positive relationship between the two methods will automatically produce a high Kappa value, because the agreement against which we (and Kappa) compare is then not a matter of pure chance. This is what happens here; the STAGES scoring system is *derived from* the method the MAP is based upon, and anything but a high Kappa is very unlikely. Therefore, the Kappa value is not wrong, but the verbal Kappa description can give a false sense of good replication, if left unchecked. The issue is not the Kappa value as such, but my knowledge of the scoring systems tells me to investigate further before I can safely say

[2] Physical Therapy & Rehabilitation Journal, Volume 85, Issue 3, March 2005

something more precise about STAGES in relation to the MAP. Again, I am investigating all this in relationship to the specific *claims* put forward by STAGES. When I asked Polissar what he knew about the relationship between the MAP and STAGES scoring system he said that he did not know anything about them (Polissar, personal conversation, 2020). In any event and more importantly, we need to factor in the quality standard, see below.

This leads me to the necessary distinction between replication, validity and validation. About this distinction, Murray says this: "*There are different interpretations of words like "replication" and "validation" within the literature, and we do not need to get bogged down in terminological nuances…."* I strongly disagree with this line of reasoning. Murray starts to relativise terminology in order to defend his own position and build a case for the strengths of the results from the replication study. From the article *Misinterpretation and Misuse of the Kappa Statistic*[3]:

> Kappa was proposed as a measure of reproducibility. When the assessment of a method focuses instead on its validity – as a surrogate for, or predictor of, the truth as estimated by a more valid reference measure – there is asymmetry between the two measurements, so indices of agreement other than Kappa may be more informative. (p. 165)

If we use Kappa as a validity proof, or convergent validity proof, we can end up in situations like this one: We take a measuring or diagnostic tool, then make something new and different, yet related, and then claim that the latter one is "validated" because of the inevitably high Kappa. In our situation somebody else can then make a variation of the STAGES scoring system, the SEGATS scoring system, do a replication study with STAGES, calculate the Kappa, and get that scoring system "validated." Next time around, the TAGESS system is developed, a new replication study with SEGATS is conducted, and the high Kappa used as a validation proof for TAGESS. Now, are we still measuring ego development? The point is that Kappa was meant primarily as a measure of replication or reproducibility, not validity or validation and hence cannot be used *alone* as a proof of validity, let alone validation.

When it comes to the importance of a minimum number of exact matches, the article above from *Physical Journal* says this: "Test the significance of kappa against a value that represents a minimum acceptable level of agreement, rather than against zero, thereby testing whether its plausible values lie above an "acceptable" threshold." In other words, we cannot use the suggested minimum 0.81 "very strong agreement" value as a threshold for our calculated Kappas since it does not consider the necessary exact agreement standard. Using the Kappa, the acceptable threshold value against which we compare must therefore be a theoretical Kappa that reflects a minimum number of exact agreements. And the minimum number of exact agreements is 85%, and no 2+ deviations, what I called "quality standard" and which reflects the reliability STAGES claims. What would the adjusted Kappa benchmark value be then? We already know that even 0.94 is not high enough (my figure 2 with less than 85% direct agreement). So, whatever that exact benchmark value may be, it is higher than 0.94 and the "very strong agreement" statement is invalidated.

I am not against Kappa at all, but here the displays of my figures 2, 3 and 6 are enough, especially with all of the more than one stage deviations. And again, I am always referring back to the specifics of the claims made in the validation article, like claiming that STAGES "retains the base

[3] American Journal of Epidemiology, Vol. 126, August 1987

and strengthen the assessments" of the original sentence completion test technology, which the MAP is based upon.

MetAware Tier

In Murrays response to my discussion on the MetAware Tier he writes that:

> Merckoll also speculates upon scoring issues specific to the MetAware level…. Merckoll ignores that fact that we have two additional published studies. … these more speculative assertions would seem to be evidence that the author is stretching to make a pre-determined point rather than offering solid evidence.

First, I do not ignore the other studies, nor do I speculate. I am referring to something I have directly observed, both in my own inventories and in conversation with several others (from areas including spiritual teachers, university professor, integral psychotherapist, and leader of integral salons). These are all qualified people; hence I can use the term anecdotal evidence. What I am merely trying to bring to Murray's attention, is that there seems to be a tendency to score some inventories prematurely at the MetAware Tier. This is what several people have observed either in themselves or in their interactions with others, based on experiential understanding of EDT and/or through the MAP (or a similar instrument). I wrote to find reasons for why this may be the case, since all my previous attempts to get Murray's attention failed. Murray can dismiss this as he wants, but I am not just speculating, and I was giving some honest arguments for further discussion. As for longitudinal studies on the MetAware Tier, they will still work as they do, even though they may include inventories that would not have qualified as such using the MAP (or a related original sentence completion instrument).

But since Murray has stated that *we have never said that "if you receive a score on a STAGES assessment it should match well with what you would get on a MAP assessment"* we can just park this discussion. It is again up to Murray how he responds to information from colleagues in the field.

If you have been scored at the MetAware Tier by STAGES, great, and I am not in a position to say that is wrong in any general way. Even more so since the definition belongs to STAGES and not me. Wherever you are is exactly where you should be. In any event, assessments like these should not be taken too seriously by being used outside of their inherently limited domain. What I can say though, as a rough guide, is that to the extent that you feel important or proud by that MetAware assessment, and especially if you need to tell everybody about it, that part of you is very likely not Construct Aware, let alone Unitive, as defined by EDT.

And finally, once again, I am not at all against the theoretical distinctions O'Fallon has brought to this late-stage territory. It is something worth honoring and for those inclined I recommend having a look at it. As the article says*:*

> In fact no one else is doing serious research on these later levels, and as far as we can tell STAGES has the most robust data set for doing research in the later levels, not to mention a

model that articulates more structure and nuance beyond Strategies/4.5, vs. alternative models.

Brief Comments on the Other STAGES Studies

According to Murray I have not paid sufficient attention to the other studies on the STAGES scoring system.

> Merckoll was aware that several other studies of STAGES were completed following the Heliyon study but decided not to read them prior to writing the critique. We wish that Merckoll had read about these studies before submitting his critique, which ignores the implications of the other studies to paint what we think is an unfairly negative picture.

I understand that it looks as an unfairly negative picture, but I merely examined STAGES own empirical data in the replication study and compared to the specifics claims made. And to bring my point home, I unpacked several implications for what that kind of data would lead to over time, and why I think they may have been wrongly used as a validation proof. This rests with the validation study, and we can look forward to new evidence when we have it. I welcome that evidence fully, and I have no problem of painting a picture that is in accordance with that evidence.

The truth is that I did have a look at other studies, and as I said in my 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*. The reason I did not spend too much time on this was that I did not want to get sidetracked into other studies until I felt my questions to the replication study were really answered. In Murrays response here I feel that is done, as I have explained above.

Having looked a little deeper into the studies, I have some specific questions to aspects of the studies I will not discuss here; that is for another time between Murray and myself. But since there are no more studies including comparison with the MAP (or similarly scored inventories) I have no basis for saying anything about that relationship. And I have no real basis for saying anything conclusive about STAGES own validity. I am not dismissing the studies, but they examine STAGES inventories, and the results from those inventories will exhibit patterns, but I cannot use those patterns to say that the inventories themselves were *accurately scoring* anything. I can't say they did not score accurately either. But either way, since STAGES does not claim to score the same as the MAP, we are over in a very different discussion than the one I wanted to have in the first place.

But what is still relevant for our discussion here is that I found no other valid studies on inter-rater reliability. By "valid" study I mean blindly co-scored outside of the training context, which I explained in my article in the section on "The STAGES Certification Process vs the Replication Study." Therefore, there are no studies that conclusively proves that the inter-rater reliability matches the accepted quality standard set by STAGES in real life situations. What Murray say about this in his response, that they "*have also published other papers on the inter-rater reliability of STAGES scoring"* is really not the case. I asked him to tell me exactly which studies I should look at (e-mail Sept 22, 2023) and got this response from Murray the same day: "*Just the one in*

the special issue. As I said, that one is based on the scoring training and I said in my note to you that we do need to look further at ongoing interrater statistics."

What the reliability should be then, is up to STAGES. Perhaps 85% is unrealistic for this type of assessment and the bar should be lowered? And should deviations of more than one stage be accepted? Much of my discussion rests on this very issue. More studies can provide answers, which is the theme for the next section.

More Empirical Studies on STAGES

Any dispute on the relationship between results from the MAP and STAGES can easily be settled simply by conducting a new and proper study. Some years ago, Susanne Cook-Greuter offered Terri O'Fallon to take 100 or so inventories scored by STAGES and have them blindly co-scored using the MAP. That was to help out O'Fallon to get more empirical information about her new scoring system. Cook-Greuter offered it to be scored by other scorers than herself, and even to do it for free. However, O'Fallon declined the offer. Later, Beena Sharma repeated the offer of co-scoring to O'Fallon, but the offer was again declined.

Early in 2020, I took initiative to conduct an ego development project among a group of fellow Zen practitioners. I spoke to Cook-Greuter who was willing to do the scoring pre and post, which was at the end and the beginning of the project, planned to last at least 18 months. I then reached out to O'Fallon, who was also very positive and forthcoming. My idea was to have each inventory blindly co-scored by the MAP and STAGES, in order to learn more about the two approaches. I then asked Cook-Greuter and Beena Sharma if it was OK to have the inventories blindly co-scored with STAGES, and they had no objection; they were both very positive. The last thing was then to get the green light from O'Fallon too, so I sent her an e-mail to let her know. However, she never responded, and I never heard anything back. The project had to go on with the MAP, and unfortunately, we never had the chance to do the comparisons with STAGES. Some participants also took the choice of having an optional Lectical assessment, so that those most interested could learn about that approach too.

In previous communication with STAGES and Tom Murray in particular, I have suggested ways for STAGES to perform a new study to get better information about their approach. Basically, it is not complicated to sample a set of inventories and have them blindly scored by more than one STAGES scorer (now fully certified), and more than one MAP (or similar) scorer. Such a sample done properly, which STAGES already know very well how to do, can give valuable information about reliability and validity for both methods, as well as the actual difference between them. I outlined some of those details to Murray (2020) and I leave this to him to act upon as he chooses.

Conclusion

I definitely welcome STAGES as a new and novel theoretical model. As for the scoring system, I am pleased to see that STAGES no longer claims to give the same results as the MAP. I suggest that STAGES comes up with their own stage descriptors for each altitude. For example, Spiral Dynamics calls the Teal altitude "FlexFlow," and EDT uses "Autonomous." I'm sure STAGES can come up with their own very good ones.

As it is now, I cannot endorse the STAGES scoring system, as there are still unresolved issues relating to the evidence I have seen. However, if new compelling evidence is put forward, examined in accordance with updated claims STAGES make about their scoring system, I will keep my promise and endorse it publicly.

I react to how STAGES relates to empirical evidence, and I find the way Murray uses statistical terminology and concepts to give credibility to STAGES problematic. I think this deserves to be discussed properly in the International Consortium of Integral Scholars, and not among psychometricians not familiar with developmental theory or integral tenets.

With more powerful AI available now, this is already changing the game of assessments. However, I believe that the kind of dialogue I have been bringing forth here will continue to have value for those interested.

I'd like to end with some words from the late Daniel P. Brown, which I've kept in my heart since I heard them. I had the great fortune to ask Brown some questions about ego development and STAGES one of the times I attended his retreats. At the end of our dialogue, I asked him how he himself saw these unfolding stages? His response was: "As rays of light."