Brief Comparison of Five Developmental Measures: the WUSCT, the SOI, the LDP, the MAP and the GLP

William R. Torbert © 2019 *
Principal, Action Inquiry Associates
Leadership Professor Emeritus, Boston College

* With gratitude for their contributions and feedback to Jennifer Garvey Berger, Elaine Herdman Barker, Terri O'Fallon, Chuck Palus, John Sabbage, and Beena Sharma.

Executive Summary

A brief comparison is offered among five related adult development measures – the Loevinger Washington University Sentence Completion Test (WUSCT), the Kegan Subject-Object Interview (SOI), the Harthill Leadership Development Profile (LDP), the Cook-Greuter Mature Assessment for Professionals Profile (MAP), and the Torbert/Herdman-Barker Global Leadership Profile (GLP).

This comparison puts special emphasis on the criteria of pragmatic and transformational validity and efficacy – that is, on the relative usefulness of the different measures in leadership development and organizational change efforts. On these grounds, the GLP and the MAP are found to have the best claims to validity and usefulness at this time.

Finally, the GLP is described as possessing several unique advantages, including: 1) the latest reliability tests of its scoring, as well as reliability testing on every single client-profile; 2) the longest tradition of empirical findings, using different methods, of the different effects of different leadership action-logics in the field; 3) the fact that it scores only through the Early Alchemical action-logic (the latest action-logic that makes an empirically-demonstrated leadership difference); and 4) its anchoring in the new Collaborative Developmental Action Inquiry (CDAI) paradigm of science.

What follows is a brief comparison among five related adult development measures – the Torbert/Herdman-Barker Global Leadership Profile (GLP), the Harthill Leadership Development Profile (LDP), the Cook-Greuter Mature Adult Profile (MAP), the Kegan Subject-Object Interview (SOI), and the Loevinger Washington University Sentence Completion Test (WUSCT). This comparison is, as you will see, not complete; rather, it puts special emphasis on the criteria of pragmatic and transformational validity and efficacy – that is, on the relative usefulness of the different measures in leadership development and organizational change efforts. (For prior peer reviewed published studies of the 45-year history of reliability and validity testing, see Torbert & different co-authors, 1987a, 1987b, 1994, 2004, 2009, 2013, available at the www.actioninquiryleadership.com website.)

Before making the comparisons among these measures, I offer a short history of my relationship to all of them. After making the comparisons, I offer a short history of the reliability and validity testing of the sentence-completion adult development measures (the WUSCT, LDP, MAP, and GLP), which are all predominantly based on sentence stems from the WUSCT.

My Relationship to the Five Measures and Particularly to the Four Sentence Completion Measures (WUSCT, LDP, MAP, GLP)

Before offering the actual comparison among the measures, it is important to clarify my relationship to these five approaches, so that the reader can appreciate both my close familiarity with all five as well as my possible bias, given my current engagement with the GLP. I have been a colleague of Bob Kegan's since 1974 (long before he and his associates created the SOI), as well as a colleague of Susanne Cook-Greuter's since 1980, and of the Harthill owners from 1992. I also corresponded repeatedly with Jane Loevinger, who created the WUSCT measure, in the early 1980s, as I and my associates adapted Loevinger's measure and language into the Leadership Development Profile, with the capacity to offer feedback to in-the-field leaders. Elaine Herdman-Barker became a certified scorer and clos colleague in the early 2000s.

I began doing empirical field and laboratory research on the relationship between people's developmental action-logic and their organizational leadership actions and effects in 1980. I have published peer-reviewed quantitative, qualitative, and action research, alone and with many colleagues on these issues ever since (most recently my June, 2013, ***Integral Review*** summary article, http://www.integral-review.org which contains the most detailed statistical analysis of the action inquiry approach to leadership development and organization transformation).

At the outset, I used Jane Loevinger's WUSCT in my adult development and organization transformation research, replacing four of the sentence stems with work-related stems that had been independently validated (her form had been originally created

for teenage girls and had no work-related stems). I engaged the Loevinger-certified scorer, Susanne Cook-Greuter, to score the sentence completion forms. I already had developed a theory of adult development that highlighted an early action-logic (the Expert) that Loevinger did not mention except as a transition point, that named all of the action-logics very differently, and that defined later action-logics very differently from Loevinger (what the GLP today names the Redefining, Transforming, and Alchemical action-logics). Cook-Greuter also diverged from Loevinger in her understanding of the later action-logics, and she eventually established the reliability and validity of a different conception and different scoring categories for the later action-logics in her 1999 Harvard doctoral dissertation (with both Kegan and myself on her committee). During the 1990s and early 2000s Cook-Greuter and I allied with Harthill Consulting in the UK to make what we then called the LDP commercially available, and we did some of the field research validating the measure during this time.

By 2005, Cook-Greuter separated from Harthill, allied with the Integral Institute, renamed her measure the MAP, with minor wording changes in some of the sentence stems from the LDP. In 2010, the Harthill owners, in a divorce proceeding, claimed the LDP intellectual property in its then current form as theirs exclusively (although they could not claim the sentence stems since they are in the public domain), and I too separated from Harthill. In 2012, Elaine Herdman-Barker (a Cook-Greuter trained scorer, as well as executive coach and consultant) and I created a slightly-different version of the instrument – the GLP (with the Loevinger and other pre-Harthill stems still making up 24 of the 30 GLP stems, and the GLP still scored by two of the same two scorers as the LDP had been).

Along with the slightly-revised instrument itself, Herdman-Barker and I also developed: 1) an entirely new set of feedback materials for business and research clients; and 2) invited 28 additional scholar-practitioners to form the Action Inquiry Fellowship. This fellowship is intended to guard the integrity and ongoing inquiry process around the paradigm, theory, research, and practice formally known as Collaborative Developmental Action Inquiry (CDAI), which embraces and interweaves 1st-, 2nd-, and 3rd-person research and practice, including the GLP as our primary 3rd-person psychometric instrument. We test the validity and reliability of the GLP, not only in the field and statistically, but also through our efforts at timely action at our AI Fellows' twice-annual three-day retreats, our collaborative research and writing, our leadership development workshops, and our client engagements.

Thus, the reliability, validity, and field testing of the WUSCT and the LDP (when Cook-Greuter and Torbert were associated with Harthill) stand behind both the MAP and the GLP, with new methods and practices emerging in recent years, as will be discussed in more detail below. First, however, I offer the following direct comparison among the five measures in terms of their validity and efficacy from a client's perspective.

Pragmatic and Validity Considerations in Choosing Among the Five Measures

The SOI. From the client's perspective, I first eliminate Kegan's SOI from further consideration – even though it is theoretically the most elegant and methodologically the most differentiated of the theories and measures. I nevertheless eliminate the SOI on pragmatic grounds because it is an interview of an hour or more, requiring further hours to analyze and score each interview. It is thus virtually impossible to administer, score, or debrief in any kind of bulk at any kind of reasonable price for large scale research, executive coaching, or consulting samples.

Nonetheless, for those trained in the SOI approach, it can be very useful for coaching executive clients and consulting to small executive teams. Moreover, Kegan-student Jennifer Garvey Berger and her partners have created the most deep-ranging global monthly conversations about developmental theory and practice through their Growth Edge Network calls. A recent [Helsing & Howell, 2014] *Journal of Management Inquiry* article illustrates both the SOI's research potential and its limitations. Bob Kegan et al.s 2016 book, *An Everyone Culture*, highlights three large, organizations that promote all their employees' development.

The WUSCT. Next, I eliminate the Loevinger WUSCT from further consideration – even though it is the original instrument from which the LDP, the MAP, and the GLP have grown… and even though a majority of the stems in each of the three more recent versions are stems from, and largely validated by, the initial WUSCT research… and even though, as well, many of the psychometric validity tests (i.e. tests of predicted correlations in relation to other measures of psychological constructs) are tests using the WUSCT, to which the LDP, MAP, and GLP refer back as partial grounds for their validity.

My three main reasons for eliminating the WUSCT from consideration are that: 1) it lacks all face validity for professionals and leaders seeking to use it for insight and increased leadership action efficacy, since it lacks sentence stems relating to the work world, as well as any history of, or client report for, offering any feedback to support leadership development; 2) it contains a great deal of evaluative language to which clients and h.r. departments object; and 3) as Cook-Greuter (1999) long ago showed, the WUSCT also lacks both a coherent theory of, and methodology for, scoring the late Transforming and Alchemical action-logics. (In particular, the GLP code book for scoring a sentence stem as Alchemical is entirely different from Loevinger's, and differs from Cook-Greuter's MAP code book in that includes all of the MAP criteria for both the Alchemical and the next MAP stage, plus an additional criterion.) This is a key point because, as soon as organizations come to appreciate the rare competitive and cooperative advantages offered by senior executives measured as operating from the Transforming and early Alchemical action-logics, instruments capable of distinguishing such persons gain recognition for their educational and practical values.

The LDP. I turn next to the LDP as the chronologically first of the three measures and processes of feedback that share the most common heritage, that are most

similar to one another, and that are most directly focused on leadership development (namely the LDP, the MAP, and the GLP). In earlier years, when Cook-Greuter and Torbert were engaged in research and scoring on the LDP, it generated some of the most well-known findings supporting the transformational efficacy of CDAI. Since 2010, however, the LDP has lost both research principals of the measure (Susanne Cook-Greuter and Torbert) and both its reliably-trained scorers. Consequently, the LDP today lacks properly trained scorers or any recent, current, or going-forward publicly available research on validity or reliability, let alone any peer-reviewed journal articles involving the current LDP. Together, these lacks make the LDP difficult to recommend.

By this process of elimination, I come to judge the MAP and the GLP as the two most practical and defensibly valid measures of leaders' developmental action-logics at present.

Comparing the MAP and the GLP

Now I turn to the similarities and differences between the MAP and the GLP. I would say that the similarities include:

a) the fact that they both benefit from Torbert and Cook-Greuter's quarter century of work in tandem on validity and reliability studies of the LDP, with Cook-Greuter and Beena Sharma now leading the MAP work and Torbert and Elaine Herdman-Barker leading the the GLP work;

b) the fact that both versions of the measure are scored by Cook-Greuter-trained scorers;

c) the fact that they both benefit from Torbert's practical, theoretical, and empirical research into individual, organizational, and scientific developmental action-logics, which is responsible for much of the scientifically-tested external validity these measures can claim (i.e. their pertinence to questions of how leaders and organizations transform in the real world);

d) the fact that both the MAP and the GLP have debriefers and executive coaches trained and authorized by Susanne Cook-Greuter (MAP) and Elaine Herdman-Barker (GLP);

e) the fact that Cook-Greuter and her MAP-trained scorers, as well as Torbert, Herdman-Barker and other Action Inquiry Fellows, are all (relatively) advanced 1st-person action researchers (thus relatively aware of the way and degree that our actions and interpretations influence the clients with whom they are working).

Differences that I see between the GLP and the MAP include:

a) the MAP report is somewhat more oriented toward psychological, theoretical, and methodological issues, whereas the GLP report is somewhat more oriented toward practicing leaders.

b) certified MAP coaches receive nine days of workshop training, whereas certified GLP coaches receive three days of workshop training and supervision with two or more GLP-scored client-leaders;

c) the MAP report is about three times longer than the GLP report;

d) there are more recently published reliability and validity statistics on the GLP scorers (see below) than on the MAP scorers;

e) reliability is checked on every single GLP score, as well as against client self-estimates; thus, reliability is tested for every client score (in a somewhat similar vein, the MAP scorers participate in regular calls to discuss interesting scoring and debriefing challenges and engage in ongoing training to maintain reliability levels);

f) the GLP has renamed the 'Individualist' and 'Strategist' action-logics 'Redefining' and 'Transforming,' responding to concerns that the former names did not convey well the leadership advantages of those two action-logics;

g) perhaps *the biggest technical and theoretical difference* between the two measures is that the MAP scoring system claims to be able to find differences that make a leadership difference beyond the Alchemist developmental action-logic; by contrast the GLP scoring system ends at "Early Alchemist" (see explanation below). In support of the GLP approach, I know of no refereed journal article, nor publicly available white paper, supporting that MAP claim.

h) perhaps *the biggest practical difference* between the MAP and the GLP occurs **after** you've taken the measure and debriefed the MAP or GLP report. While the debriefing and coaching with regard to both measures is excellent, the GLP follow-up includes the whole arena of CDAI practices tested in many organizations. The CDAI approach includes not only the personal developmental action-logics, but also the organizational development action-logics, the scientific action-logics, the three types of feedback, and eight different types of power that can be interwoven in action. Thus, CDAI explicitly integrates 1st-, 2nd-, and 3rd-person research in the midst of practice, providing researcher/practitioners with a complex set of perspectives and the post-perspectival awareness relevant to taking timely action.

A Brief History of the Cumulative Reliability and Validity Testing of the Sentence Completion Measures

This section offers an overview of the history of reliability and validity testing of the four sentence completion measures, all of which share about 80% of the same sentence stems.

In the 1960s, when I was learning social science at Yale, the most advanced and authoritative methodological wisdom by those working within the Empirical Positivist paradigm (Popper, Campbell and Stanley, etc.)… Was that, if you wanted to develop and use a new measure that relied on expert scoring (like the SOI, WUSCT, LDP, MAP, GLP, or StAGES), you first needed to develop high reliability among scorers (.80 at minimum; over .90 impressive). Otherwise, its scoring would be relatively arbitrary and any attempts at correlating the measure in predictable ways with other psychological or sociological measures or real world outcomes could not help either failing outright or, if appearing significant, being spurious.

If you succeeded in the first task (achieving reliability), you would next establish "internal validity" or "psychometric validity" which, broadly but incompletely, means testing it against other measures that have already proven their worth in the social science community. Loevinger and her students' work with the WUSCT addressed itself largely to these two issues of reliability and internal validity. Thus, for example, Loevinger trained scorers, like Susanne Cook-Greuter, to high levels of reliability, and tested the WUSCT against many other measures, such as Rokeach's Closed-Open Values measure, accurately predicting a high correlation between later developmental action-logics and Openness. Years later, while Cook-Greuter and I were researching and using the LDP, she conducted most of the training of two additional scorers, who attained a 84% perfect agreement reliability score. Both these scorers and I continued working with the LDP for several years, before creating the GLP.

After the stages of internal and psychometric validity testing, according to methodological opinion, you explored the measure's "external validity," which, again broadly and again incompletely, means testing the measure against some predicted real-world outcomes. This is the type of testing my colleagues and I have been most interested in doing over the past 30+ years in relation to the LDP/GLP family of instruments. Statistically significant results of external validity testing not only reinforce the validity of the instrument itself, but also provide new substantive social science knowledge. For example, we have found very powerful correlations (accounting for more than half the variance in the outcomes) in terms of the organizational action-logic necessary before an organization systematically supports leadership development at work (Torbert & Fisher, 1992). Another finding that accounts for more than half the variance shows which leadership action-logics are necessary to reliably succeed in generating organizational transformation (Rooke & Torbert, 1998; Torbert, 2013). More specifically, only those CEOs and lead consultants who have measured at the Transforming or early Alchemical action-logics reliably generate positive organizational transformation (leading to larger market share, profits, and enhanced reputation).

Once a measure has satisfied reliability, internal-validity, and external-validity criteria, it and/or its scorers need only satisfy external-validity tests henceforward, since no measure can repeatedly show statistically-significant, externally-valid results unless it is being scored reliably and actually measures the variable it claims to measure. The

external validity tests become increasingly persuasive as one uses additional methods and research designs (e.g. laboratory experiments, field action research, interviews, non-obtrusive measures, etc.), as has been the case in the external validity testing related to the LDP/GLP family of measures (Torbert, 1994).

Once scorers of a measure have attained reliability as determined by traditional reliability tests (where both scorers separately score the same set of items and their level of agreement is assessed), the best way to maintain and increase reliability and accuracy on a measure by measure basis is to conduct a reliability test on the scoring of each measure, by having a second scorer review the first. This kind of reliability test has been conducted twice in recent years on the GLP scorers. In a 2009 review of 805 measures, each of which could have been scored at 13 different levels (e.g. Early Diplomatic, Diplomatic, Late Diplomatic, etc.). The result showed a .96 Pearson correlation between the two scorers, with perfect agreement in 72% of the cases, with a 1/3 action-logic disagreement in 22% of the cases, and with only one case of a disagreement larger than one full action-logic. The cases of disagreement led to negotiated agreements prior to feedback to clients, and these agreements were presumably more accurate in most cases than the original score. Thus, this process not only served as a reliability test, but also made the results offered to clients more accurate.

In early 2016, a stratified sample of the 78 most recent GLP sentence completion forms from 2015 (10 Expert, 20 Achiever, 20 Redefining, 20 Transforming, and 8 Early Alchemical) were reviewed for reliability between the scorers, not only in terms of protocol scores, but also of item scores within each protocol. This study found perfect agreement on the protocol score in 94% of the cases and only a 1/3 action-logic disagreement in the other 6% (remembering that these disagreements were abjudicated prior to release of the results to clients). In terms of item scores, there was 98% perfect agreement between the two scorers, only 4 cases in 2340 stem scores of a two-action-logic disagreement, and no cases of a larger disagreement. When one compares these results to the ones seven years previously by the same two GLP scorers, one sees a 22% increase in perfect agreement. This increase in agreement presumably occurs at least in part because of the continuing, measure by measure comparisons of scores between the scorers throughout the years.

Some argue that a short reliability test where each rater rates the same ten sentence completions without knowledge of how the other is rating them is a ‘stiffer’ test. But the glaring, undiscussed weakness of such reliability tests, for which scorers can specially prepare and take under optimal conditions of rest and alertness, is that you can never tell whether they in fact generalize to the scorers’ day-to-day scoring.

A second argument against this kind of reliability test is that the second scorer will be motivated to agree with the first in order to achieve high percentages of agreement. However, the two scorers were not aware that their work would later be subject to a reliability test. Moreover, the second scorer has two much more operative motivations for honestly disagreeing with the first scorer when that is the case: 1) if clients’ receive systematically undefensible scores not useful in their further development, GLP work will dry up altogether; and 2) if GLPs are not accurately scored, they will be of no use in training future scorers how to score.

In addition to this novel commitment to measure by measure reliability testing between scorers, every 400-word profile commentary is also reviewed, and every client/research-participant who takes the GLP is introduced to a systematic self-assessment process of their own action-logic prior to receiving feedback about their center-of-gravity action-logic from the GLP. In a 2015 test of 66 cases, clients' estimates differed from the GLP analysis in only 6 (or 9%) of the cases. In 4 of these cases, discussion with the debriefer led to client agreement with the GLP score, usually because the client realized that the conditions they blamed for their score (e.g. that they were rushed or tired) were their usual conditions at work and therefore reflected their general work action-logic. In the other 2 cases, the GLP debriefer came to agree with the participants' self-estimate, because English was not the respondents' first language and affected their written. Thus, in 94% of the cases studied, the client accepted the validity of the GLP analysis. This process of case-by-case exploration with research participants is, simultaneously, a new form of reliability testing and a new form of external validity testing.

In late 2016, two new GLP scorers completed their training with Elaine Herdman Barker. A test of reliability between the new scorers' ratings of each of the 30 sentence stem responses on 30 protocols (n=900) and Herdman-Barker's scores (of which the new scorers were unaware) show the levels of precise agreement at 87.1% and 87.4%, with a disagreement of two levels occurring in less than 1% of the cases. In 2018, another scorer completed training, showing 89.9% precise agreement on stem scores, as well as 70% perfect agreement on whole-protocol ratings, with the other 30% within 1/3 of an action-logic (e.g. Achiever and Late Achiever). This results in a Spearman correlation of .98.

A Construct Validity Test. In a different domain, a 2009 test of the validity of the sentence completion measure involved predicting a difference that should theoretically exist between early action-logic and late action-logic protocols, and then testing whether that difference exists empirically. Using 891 measures, a cluster analysis test of construct validity was conducted to assess whether the theoretically-predicted difference in cognitive complexity between Conventional action-logics (up through late-Achiever) and Post-Conventional action-logics (starting at early-Redefining) could be found in a statistically different clustering of factors in the two sub-samples. The prediction was that Conventional thinking involves sharply distinct classes of phenomena, whereas Post-Conventional thinking involves more mutable conceptual boundaries. We analyzed the underlying pattern of two separate sub-samples: 1) 830 protocols rated overall as 'Conventional'; and 2) 61 protocols rated overall as 'Post-conventional.'

We found a remarkable qualitative difference between the Conventional and Post-Conventional cluster analyses, immediately observable in the two figures in Livne-Tarandach and Torbert (*Integral Review*, 2009). For the Conventional action-logics, stem-response ratings load on eight distinct factors, each of the eight nodes indicating a similar pattern of answers and scores. For example, stems 3, 17, 20, 22, 24 that make up cluster 6 reflect the high correlation in scores assigned to this set of stem-responses across the different respondents. Overall, this cluster analysis of the factors, or overarching themes, that emerge when analyzing Conventional protocols is itself quite

conventional statistically: distinct clusters or factors show up, with different sentence stems associated with each distinct factor. The factors are, in short, conceptually distinct.

In contrast, we found a strikingly different pattern emerging from Post-Conventional profiles. For the Post-conventional action-logics (Redefining and later), stem-responses loaded on 11 factors, but loadings were not confined to one factor per stem. More than half (52%) of the stems loaded on two factors or more (9 stems loaded on 2 factors, 7 loaded on 3 factors, and 3 loaded on 4 factors).

The stably-focused Conventional factor loadings represent a relatively simple mental map, with Aristotelian-ly distinct, independent, lasting categories ("nothing can be both A and not-A"), as one would theoretically expect of action-logics up through the Conventional. In contrast, the complexity of the Post-Conventional sets of loadings suggest that Post-Conventionals hold a systems-oriented, inter-independent, wedding-of-opposites 'living' mental mapping process.

Plato's two distinctive images for the nature of thought in the *Thaeatetus* – as either 'marks on a wax tablet' of the mind, or 'birds flying about in an aviary' of the mind – seem remarkably apt as metaphorical summaries of the statistical difference between Conventional and Post-Conventional thought.

The ability of the post-WUSCT sentence completion measures to make such distinctions, and the ability of CDAI organization transformation practices to support both individual and organizational development again confirm the credibility and efficacy of the GLP, when used within the CDAI paradigm of research and practice.

For Further Reference

Torbert, Cook-Greuter & Associates' 2004 book *Action Inquiry: The Secret of Timely and Transforming Leadership* is the most accessible and complete practitioner's guide to CDAI, and also includes a review of scientific literature in its "Concluding Scientific Postscript." Together, Torbert's 1973 book *Learning from Experience: Toward Consciousness* and his 1991 book *The Power of Balance: Transforming Self, Society and Scientific Inquiry* convey the theoretical and existential foundations of "living inquiry." Cook-Greuter's 1999 Harvard dissertation, *Postautonomous Ego Development*, offers and methodologically defends her different construal from Loevinger of the later action-logics. Livne-Tarandach & Torbert's 2009 *Integral Review* article and Torbert's 2013 *Integral Review* article "Listening into the Dark: An essay testing the validity and efficacy of CDAI for describing and encouraging transformations of self, society, and scientific inquiry" offer the most complete exploration of the grounds for trusting and continuing to test the validity of CDAI and are available at http://integral-review.org.

Other Torbert publications referred to in this paper can be found at (and downloaded from) www.actioninquiryleadership.com In addition, seven of the most recent doctoral dissertations that are grounded in, and that innovatively extend, CDAI theory, practice, and method are authored by CDAI Fellows Aftab Erfan, Ed Kelly, David McCallum SJ, Cara Miller, Aliki Nicolaides, Shakiyla Smith, and Karen Yeyinmen (McCallum, Nicolaides, Smith, and Yeyinmen have written doctoral dissertations based on empirical research using the GLP). Other Action Inquiry Fellows who both practice and publish in the area of CDAI include Hilary Bradbury-Huang (Editor-in-Chief, *Action Research* journal and *Handbook of Action Research*), Erica Foldy (Professor, NYU Wagner School of Public Service), Jenny Rudolph (Director, Medical Simulation Center, Harvard Medical School), Steve Taylor (WPI Management Professor and Editor-in-chief, *Organizational Aesthetics*), and Nancy Wallis (Pitzer College). Again, some of their work can be found in the CDAI Cyber Library at www.actioninquiryleadership.com).

To access and use the GLP for consulting, coaching, or research purposes, please see www.gla.global