Reliability and Validity Tests of the Harthill Leadership Development Profile in the Context of *Developmental Action Inquiry* Theory, Practice and Method

William R. Torbert, Reut Livne-Tarandach

Abstract: In this paper, we describe how the Harthill Leadership Development Profile (LDP), a language-based instrument has evolved from Jane Loevinger's Washington University Sentence Completion Test (WUSCT), and has been redesigned to assess and offer feedback about adults' action logics in work or educational settings, in the context of Developmental Action Inquiry (DAI) theory, practice, and method (Torbert, 1972, 1976, 1987, 1991; Torbert & Associates, 2004).

Next, we challenge a recent critique of the LDP as a soft measure unsupported by published, quantitative psychometric reliability and validity studies (Stein & Heikkinen, 2009) and present both previously unpublished and previously published-but-not-aggregated studies illustrating Harthill LDP as a well-calibrated measure of adult ego-development.

Because the DAI approach to social inquiry and social practice invites us all to interweave first-, second-, and third-person inquiry and everyday action, the validity studies reported tend to concern field-based experiments seeking to generate developmentally transforming change in adults, including the researchers and/or interventionists, as well as in the organizations in which they participate.

In our conclusion, we briefly consider what a social science and a social practice based on the developmentally late action-logics will look like, once social science is recognized as embracing, not just 3rd-person empirical positivist research "on" subjects, but also 1st-, 2nd-, and 3rd-person research and action *with* co-participants in live settings.

Keywords: Action-logic, developmental action inquiry (DAI), Harthill leadership development profile (LDP), reliability, validity.

Developmental Action Inquiry: Theory, Practice and Method

Developmental Action Inquiry (DAI) is an approach to social science and to personal and social life that posits a sequential series of increasingly complex, inclusive, and mutually-transforming action-logics through which individual persons, conversations, relationships, organizations, and scientific paradigms may evolve as they intertwinedly act and inquire.

Drawing on the first-person action research of the Gurdjieff Work tradition (Ouspensky, 1949), as well as on the second-person "action science" of Chris Argyris (Argyris, Putnam & Smith, 1985), and on the third-person developmental theorizing of Erikson (1959), Torbert (1972, 1976, 1987, 1991, 2000a, 2000b; Torbert & Associates, 2004) has demonstrated how to interweave first-, second, and third-person research methods with first-, second-, and third-person practices to study the past, the present, and the future in ways that generate single-, double-, and triple-loop feedback (Chandler & Torbert, 2003; Starr & Torbert, 2005) that leads to increasingly timely action and inquiry (Torbert & Martinez, forthcoming). We will not summarize the DAI approach any further here, but the reader will find it illustrated in the validity studies described below, and the approach is reviewed in Herdman-Barker, Livne-Tarandach, McCallum, Nicolaides, & Torbert (2010).

In 1978, Torbert chose to begin working with, and adapting, Loevinger's Washington University Sentence Completion Test (WUSCT) as the third-person empirical measure of development that most closely tracked DAI theory at that time. We now turn to the question of how the WUSCT was gradually adapted to become the Harthill Leadership Development Profile (LDP).

The Development and Evolution of the Harthill LDP

The Harthill LDP (LDP) is a language-based instrument designed to assess adult developmental action-logics as part of the third-person empirical testing of DAI propositions. The current version of LDP includes thirty-six open-ended sentence stems that, upon completion by participants into sentences, enable one to measure adult action-logics. The HLDP evolved out of Loevinger's WUSCT (Loevinger, Wessler, and Redmore, 1970), and now includes 24 of Loevinger's original items along with 12 new items designed to improve Loevinger's original test in a number of ways.

First, in order to move away from the original focus of the WUSCT on women's issues and thereby expand the LDP's generalizability, the current version of the HLDP omits a number of gender-based items such as "The worst thing about being a woman...", "A good mother...", and "For a woman a career is...". Second, to improve the tool's face validity and usefulness as a work-related, leadership development instrument, the current version includes work-related stems, such as "A good boss…"; "A person who steps out of line at work…"; "I am powerful…"; "When it comes to organizing my time…"; and "Teams…"). As will be detailed below, the responses to the new stems correlate better with an individual's overall profile rating than responses to the former stems did, thus improving the overall reliability of the measure. Together, the additional stems incorporated in the LDP introduce a meaningful and unique contribution because they expand the scope of generalizibility of the LDP to better map management oriented themes that are at the heart of developmental implementations in organizational contexts.

In addition to gradually substituting new stems for previous ones, many other changes, described in greater detail in Herdman-Barker & Torbert (2010), have been made in the transition from the WUSCT to the LDP. The names of the developmental action-logics have been changed, along with key characteristics, to make them less evaluative, more descriptive, and more useful as feedback to respondents. DAI theory posited an action-logic between Loevinger's 3 and 4, and we found that the highest proportion of most of our samples scored at this in-between point, which we named Expert and defined. DAI theory also defined the latest (and rarest) action-logics quite differently from Loevinger, naming them Alchemist and Ironist. Cook-Greuter (1999) first re-defined and successfully tested new scoring procedures for these action-logics, and Herdman-Barker and Torbert (2010) later further redefined the scoring procedure and performed the construct validity test described toward the end of this paper.

Today, the names of the DAI personal action-logics, in developmental order, are: Opportunist, Diplomat, Expert, Achiever, Individualist, Strategist, Alchemist and Ironist.

History of Reliability Testing of the WUSCT and the Harthill LDP

The 2004 Torbert & Associates book *Action Inquiry: The Secret of Timely and Transforming Leadership* (Berrett-Koehler, San Francisco) contains an 18-page appendix on the history of the Harthill LDP up to that point. That appendix contains references to a large number of quantitative studies particularly testing the reliability of Loevinger's WUSCT, from which the Harthill LDP has evolved.

We will not repeat that material here, except to let the reader know three related points. First, our own earliest attempts to do developmental research in the 1980s used the WUSCT, relying on the extensive reliability and validity findings about the WUSCT at that time, and working with WUSCT scorer Susanne Cook-Greuter to test whether respondents' WUSCT ratings corresponded to the predictions of Developmental Action Inquiry theory for how managers and other professional measured at different action-logics performed. As will be reported in the validity testing section, we repeatedly found highly significant correlations accounting for unusually high percentages of the variance in criterion variables.

Second, Susanne Cook-Greuter was a Loevinger-authorized high-reliability scorer of the WUSCT in 1980, and she was the primary scorer of all the research reported in this article until 2005. During the 1990s Cook-Greuter redefined the later action-logics and their scoring procedures, showing high reliability results. Based on 60 protocols, she and another rater trained in the new scoring procedures achieved a Pearson correlation of .95, beyond the .0001 level of significance (Cook-Greuter, 1999, p. 115).

Third, in the early 2000s, Cook-Greuter trained Elaine Herdman-Barker as a high-reliability rater of the evolving measure. Herdman-Barker is now Harthill's lead rater and executive coach using the LDP in a constructivist developmental approach to coaching with other colleagues in Harthill.

Since 2005, Harthill has conducted a further series of six reliability tests to date among the Harthill LDP's current raters. First, we tested the extent to which the set of 36 stem-responses used in LDP can be aggregated to one score reflecting the center-of-gravity profile action-logic. To do so we tested the internal consistency by calculating the Cronbach's alpha score for the 36 response ratings. Cronbach's alpha values range from zero to one, one indicates perfect internal consistency and zero indicates no internal consistency. In general, Cronbach's alpha values higher than .8 are considered satisfactory as indicators of internal consistency (Schutt, 2004). Our analysis of the internal consistency of LDP was built on 891 distinct profiles, scored between 2005-2008 and generated Cronbach's alpha of .906, a relatively high value indicating good internal consistency that justifies the aggregation of stems into one score reflecting a single action logic.

Moreover, as the LDP includes six new stems, we assessed the extent to which these stems contribute rather than dilute the LDP assessment. To do so we explored the correlation between the six new added stems and the final profile scores and found that these correlations were relatively high (.86-.89), and that these were equal to or higher than the correlation found in previously used stems. This suggests that the newly introduced stems have added to the reliability of the measure and therefore seem to be adding to, not subtracting from the reliability of total profile scores.

Furthermore, to assure that no order effect influenced the consistency of the scale we conducted two separate analyses. First, we calculated the average individual-response ratings for each stem to the total-protocol-rating correlations for every one of the 36 stems in our 891 sample of profiles. These correlations ranged from .82 - .89 with an average of .86 across all stems. As none of the stems showed dramatically low correlation and as no particular pattern of correlation size order was evident in this set, we suspected that an order effect is unlikely. Yet, to rule out the possibility that the gaps between stems scores and profile ratings may have been influenced by stems order within the LDP sentence completion form, we used a linear regression building on 891 profiles and found that stem order had a significant effect but that the size of its effect explained less that 5 % of the variance in the gaps between stems and profiles scores. This suggests that while order effect may have a statistically significant impact, the sheer size of its impact is negligible due to its low practical significance (Peterson, 2008).

Another dimension of reliability testing concerns the assessment of inter-rater reliability. As the majority of stems included in the current version of the LDP are built on prior validated work by Loevinger and Cook-Greuter, and since Herdman-Barker and Cook-Greuter had attained a documented inter-rater reliability of .69 perfect matches and .90 within one level (when there were seven different levels), we have explored the extent to which the six newest business-related stems introduced to this version yield consistent scores. An authoritative test of reliability is appropriate only after scoring manuals for all the new stems are completed, and the scorers have trained with them. Nonetheless, in 2005 we completed a preliminary reliability test between Cook-Greuter and Herdman-Barker that included blind rating of the six new stems of 20 LDP profiles and found .74 agreement within one level (again, with 7 different levels), illustrating an encouraging reliability for an initial test, though it will be important to show a correlation above .80 in the post-manual test with different scorers.

More recently, another inter-rater reliability test has been conducted, assessing inter-rater agreement between final profile scores across 805 distinct profiles that could have been scored across 13 levels of scoring (from Expert through Alchemist, with three possible levels at each action-logic [e.g., Early Achiever, Achiever, Late Achiever]. The test shows .961 Pearson correlation between the two raters, with perfect matches for 72% of the profiles and agreement within one part-stage in another 22%. Hence, there is either perfect agreement or agreement within one part stage in 94% of the cases. There was a difference of more than one full action-logic between the two raters in only one case of the 805 profiles scored. Moreover, it is important to note that any difference in the two scores was adjudicated prior to sending respondents their feedback packages. It is also important to note that what is different about this last inter-rater reliability test is that, as the reader may imagine, both raters did not separately rate all 805 sentence completion forms. Rather, one rater scored them all, and the second (more experienced) rater then reviewed the scoring, making changes on individual sentence stem scores and on total protocol rating scores whenever changes appeared warranted.

At first, this procedure may strike the empirically-positivistic, 3rd-person-social-science reader as fundamentally inadequate. Isn't it obvious that the second rater may be subconsciously influenced by the pre-existing score to agree with it? Especially, given the aim of any inter-rater reliability test to achieve the highest agreement possible? And certainly, these questions lead directly to the importance of conducting a traditional inter-rater reliability test when the scoring manuals for the new sentence stems become available. However, consider the following arguments in favor of taking this test seriously as one among several reliability tests.

In orthodox inter-rater reliability tests, both raters know they are involved in a test and are therefore likely to try especially hard to score carefully, making it likely that the test overstates the reliability one would find if both unknowingly scored the same protocols and these scores were later tested for reliability. In contrast, the raters in this case did not imagine they were engaged in a reliability test at all (the test was only conceived post hoc). Thus, the second rater is consciously motivated to make "corrections" whenever warranted (not "as rarely as possible"). And this for three reasons: (a) for the client's sake (to give him or her the most accurate reading possible, in order to best support their further development); (b) for the sake of the continuing training of the second rater; and (c) for the sake of a productive and possibly mutually-transforming dialogue between the raters. Moreover, the accuracy rate in the final reported scores is even higher than .96 because of the change in the score once a discrepancy has been noted. (Of course, the change itself will have been a "mistake" in some small proportion of the cases, but presumably in less than half, since it is the more experienced rater who is making the changes.)

Put differently, the statistics in this inter-rater reliability test are not statistics from a small sample of all the protocols used in one research project or another, but rather represent reliability testing done on a high percentage of the entire population measured. To be more precise, this testing has been performed on all the 2006-2008 Harthill post-conventional protocols (Individualist and later, which are the most difficult to score reliably), as well as on a sample of the earlier action-logic protocols, which represents continuing oversight and feedback to the less experienced rater, even though he has already attained greater than .85 reliability with the lead rater on these.

History of Validity Testing of the WUSCT and the Harthill LDP

As in the case of our review of reliability testing, we will pass briefly over the early history of validity testing of the WUSCT because this history has been so carefully parsed previously (Torbert & Associates, 2004, *Appendix: Concluding Scientific Postscript*; the entire 4th issue of the 1993 *Psychological Inquiry*; Westenberg et al. (Eds.), 1998). As with most psychological instruments, most of the early validity testing of the WUSCT addressed convergent and divergent validity assessing the relation between WUSCT and other psychological measures. For example, a number of personality characteristics have been found to be appropriately associated with predicted developmental action-logics. These include conscientiousness, trust, tolerance, interpersonal sensitivity, psychological mindedness, creativity, and rule-boundedness (in this last case, a curvilinear relationship, lower in pre- and post-conventional action-logics, higher at conventional action-logics)(Kohlberg 1963, 1964; Lorr & Manning, 1978; Vaillant & McCullough, 1987).

Since the divergent and convergent validity of the WUSCT as the foundation platform for the Harthill LDP had at that time been unusually well established, and because all of our work grew out of the context of Developmental Action Inquiry theory, practice, and method, wherein first-, second-, and third-person research and practice interweave with one another (Chandler & Torbert, 2003; Torbert, 2000a, 2000b; Torbert & Associates, 2004), we believed that our greatest contribution to the credibility of the instrument could be made by demonstrating the construct validity of the later action-logics, as well as the criterion-related validity of the instrument as a whole in field experiments where the intent was to support developmental transformation of individuals and organizations. A recent review of the developmental leadership literature (McCauley et al, 2006) highlights the unique contribution this body of work offers. In other words, our interest in using the LDP has been in pragmatic and theoretical questions of change in the real world, such as: Do managers at different developmental action-logics act differently? What kind of educational organizational structures and strategies support, not just incremental, but also transformational change in students or employees or leaders or oneself? Do leaders at later action-logics support organizational transformation more reliably than leaders at earlier action-logics?

Our use of the instrument began in 1980 when one of us was serving as graduate dean of a management school that was pioneering an “action-oriented” MBA program and that received a research grant to study the efficacy of the new program. Thus, for three years running, all members of the entering and exiting fulltime, two-year MBA cohorts (about 96 per year) were offered the opportunity to participate in research that included the WUSCT in its initial period of transformation toward becoming the Harthill LDP. (At this point, the WUSCT was modified by four managerial stems (e.g., “A good boss…”) that had been independently validated (Molloy, 1978), and language that was more managerial and less evaluative (thus, with greater face validity) had been developed for offering feedback. Participation was voluntary, and participants were offered feedback on their scores, as well as the opportunity to participate in a related laboratory experiment, to be described shortly. Nine of 288 chose not to participate. This research eventuated in four distinct studies with implications for criterion validity in general and the predictive validity of the LDP more specifically.

Study 1

The first confirmation we received of the criterion validity of the WUSCT / LDP) occurred via the unobtrusive measure (Webb et al, 1966) of whether students in fact asked for, and followed through by appearing at the appointed time to receive, feedback on the results of their performance on the measure (Merron & Torbert, 1984; Torbert, 1994). None of those measured Diplomat (4) sought feedback, and only a small minority of those measured Expert (5) did so. A bare majority of those measured Achiever (6) sought feedback, whereas a large majority of those measured as Individualist (7) or later did so. In short, a larger proportion at each later developmental action-logic asked for feedback, a perfect 1.0 correlation on a Spearman Rank Order test, confirming the theoretical claim that later action-logics are increasingly open to (and increasingly seek out) single-, double-, and triple-loop feedback and learning. Understood this way, this finding supports both construct and criterion validity. It is worth noting that other factors such as age, gender, years of work experience did not explain the difference in behaviors we have documented.

There were some significant qualitative indicators of the validity of the measure as well during this unobtrusive test. One was that those who measured as in between the last conventional action-logic and the first post-conventional action-logic at that time (known only by the number 4/5, between what Loevinger called 4/Conscientious and 5/ Autonomous, and what Kegan ([1982)] called 4/Institutional and 5/Interindividual) were the only ones who participated in a feedback session who took the initiative to set a second feedback session. finding the new theory through which they were learning in a new way about who they were critically helpful in resolving the relativistic confusions in which they found themselves (this is the action-logic now denoted as “7/ Individualistic/ Pluralistic/ Relativistic/ Green”).

A second, equally interesting, qualitative discovery was the action-logic of the only three persons who chose to come to a feedback session who became angry at what the measure indicated about them. For example, one of the three began to tear up all the research sheets on the Dean’s meeting table until the Dean prevailed upon her sense of principle by shouting that she was not tearing up just her own research results, but those of others who deserved equal opportunity to tear up their own – a speech act that caused her to desist. All three of these ‘angry’ persons were among the four who came to a feedback session who were scored at the Expert action-logic, between the dependent Diplomat action-logic and the independent Achiever action-logic. One of the Expert action-logic’s defining characteristics is counter-dependence.

Study 2

The laboratory experiment we invited MBA students and alumni to participate in during the mid-1980s involved taking a three-hour ‘Consolidated Fund In-Basket Test’ designed by the Educational Testing Service (Merron, Fisher & Torbert, 1987). This test positions the respondents as directors of a community fund, appointed in mid-campaign because the prior director has been disabled in an automobile accident. The new director must deal with 34 in-basket items from staff, volunteers, and board members on a Sunday afternoon. Forty-nine MBA students and alumni volunteered to participate in this study, which, like the previous one, promised substantial feedback (not just written, but also in person) about their own and others’ results. The sample consisted of 29 men and 20 women, whose average age was 31, all but two Caucasian, and all of whom had held, or were currently holding, managerial positions in a variety of organizations.

None of the demographic variables (gender, age, or years of fulltime work experience) explained differences in performance. Indeed, at first, *when quality of performance was measured by executives hired by ETS to rate relative efficacy*, it seemed that developmental action-logic as measured by the WUSCT/LDP didn’t explain differences in performance either. However, when independent, trained raters measured the number of *second-order responses* (i.e., occasions when the director asked for or offered double-loop feedback in his or her notes), as well as the number of *collaborative actions* (i.e., when the director invited the other to influence the outcome), they found statistically significant differences between the respondents measured at conventional action-logics (5/Expert and 6/Achiever) and those measured at post-conventional action-logics (7/Individualist and 8/Strategist).

In greater detail, inter-rater reliability was 92% for the *second-order* variable and 85% for the *collaborative-action* variable. A chi square test showed that the relationship between post-conventional action-logic and *second-order responses* was statistically significant beyond the .01 level; and the relationship between post-conventional action-logic and *collaborative-action* was marginally significant at the .09 level. Two further confirming aspects of the findings are: (a) that, while the largest increase in *second-order* responses occurred between the Achiever and Individualist action-logics, the largest increase in *collaborative-action* responses occurred between the Individualist and Strategist action-logics; and (b) that there were only half as many respondents measured at Strategist as at any of the other three action-logics. Therefore, the marginal rather than robust significance of the second finding may be due to the under-representation of Strategists in the sample.

This study contributes to both construct and criterion validity in that developmental theory would predict such differences in action-complexity between conventional and post-conventional action-logic leaders.

Study 3

In the much larger ongoing field experiment in the Boston College MBA program, two whole classes (n=193) were profiled at the outset and conclusion of their program (Torbert & Fisher, 1992). In spite of the variety of challenges and supports for transformational learning embedded in the organization's regular functioning during the first year of the program (which included rotating team leadership roles, 360 assessments, developmental coaching, action consulting projects with local businesses, etc.), only 3 of the 177 who enacted the many developmental instructional technologies during their first 11 months transformed a full developmental action-logic after the full 22 months, according to the LDP. *By contrast, 15 out of 16* of those organizational participants who did all the above *and also volunteered for and won* a non-remunerative consulting role with new teams, during their 2nd 11-month participation (including depth-clinical training and processing on a weekly basis), *did transform a full developmental action-logic* (see details of this process in Torbert, 1991, ch. 4). (It is noteworthy that, since the participants who became team consultants on average rated at later action-logics than the rest of the participants at the outset of the process, the statistical effect of regression toward the mean was operating *against* the direction of the actual findings on the after test.

This finding accounts for an unusually high proportion of the variance (81%, Goodman & Kruskall's tau) of the participants in the field experiment who are measured as having experienced a transformation of action-logic. The overall results validate the transformational efficacy of a late-action-logic type of intrinsically educational organizing (called 'liberating disciplines' in DAI theory [Torbert, 1991]) that highlights the complementary roles, in generating adult development, of: (a) practical demands of real business clients; (b) depth mutual inquiry practices; and (c) voluntary commitment to one's own development.

Study 4

A field study of 16 MBA project teams (Torbert, 1987) showed that those teams (5 of the 16) with one or more members measured at the Strategist/8 action-logic outperformed teams with no one measured Strategist in three ways: in terms of grades on the two course projects, in terms of members' perceptions of efficient time-use, and in terms of members' perception of within-group support for own learning. (No one knew the students' LDP scores at the time when the outcome data were collected.)

DAI and the LDP as Predictors of CEO Performance in Supporting Organizational Transformations

In the 1990s, Torbert and his associates explored whether Developmental Action Inquiry theory and practice and the validity of the Harthill LDP extended beyond the educational world to the world of business organizations in dynamic markets. They consulted and conducted action research in numerous for-profit and competitive not-for-profit organizations that wished to transform their strategies, leadership cultures, and ways of doing business (Fisher & Torbert, 1995; Rooke & Torbert, 1998; 2005; Hartwell & Torbert, 1999a, 1999b; Torbert & Associates, 2004).

In ten of these organizations, the CEOs all participated in taking (and receiving feedback on) the LDP. In five of the ten cases, the CEOs scored as Strategists. In the other five cases, two CEOs scored as Achievers, two as Experts, and one as a Diplomat. There were four different lead consultants in the ten cases, three of whom scored as Strategists and one of whom scored as Alchemist. The four consultants worked in different combinations with the ten organizations for unusually long periods – an average of 4.2 years. Business and reputational measures showed that seven of the ten organizations improved dramatically during the intervention/studies, while the other three declined either mildly or dramatically. Based on the thick descriptions of the individual cases, three raters achieved perfect (1.0) reliability in determining whether an organization qualitatively transformed into a later action-logic, or did not change at all, or regressed (in one case, one rater differed over how many transformations had occurred in one company).

All five of the Strategist CEOs succeeded, with the help of the consultants, in leading their organizations through one or more organizational transformation. Only two of the other five organizations succeeded in transforming, and the lead consultant in those two cases was the one scored as Alchemist. Two showed no significant change, and the organization led by the Diplomat CEO regressed.

If one adds the HLDP scores of each CEO/Lead Consultant duo, then the combined influence of their action-logics accounts for 59% of the variance (at a .01 level, on the Spearman Rank Order test) in whether the organization succeeds in transforming. (Cohen (1983) classified a large effect size as one that accounts for 25% of the variance in a correlation test (that is, *r*=.50). A test that accounts for 59% of the variance, as this one did, represents an unusually robust empirical finding.) Thus, this study seems to support the predictive validity of the LDP and the theoretical proposition that only persons who transform to the Strategist action-logic or beyond reach the capacity to reliably support organizational transformation (Torbert & Associates, 2004).

Critical readers will, however, have many questions about possible threats to the validity of this finding, so let us examine the study more closely and consider these questions.

Testing the Third-Person, Internal and External Validity of the Ten-Organization Study

Validity criteria that test the third-person generalizability of empirical findings "after-the-fact" are enumerated and described relatively exhaustively by Cook and Campbell (1979). Their conceptualization of validity has two general components, *internal validity* and *external validity*, defined as follows:

> Internal validity refers to the approximate validity with which we infer that a relationship between two variables is causal or that the absence of a relationship implies the absence of cause.
>
> External validity refers to the approximate validity with which we can infer that the presumed causal relationship can be generalized to and across alternate measures of the cause and effect and across different types of persons, settings, and times. (Cook & Campbell, 1979, p. 37)

Cook and Campbell (1979) list 19 different potential threats to internal validity and 13 different threats to external validity. They suggest that researchers focus on the threats most likely to have a significant effect on the validity of their work. In Cook and Campbell's terms, this study is best described as a *nonequivalent control group quasi-experimental design*, whose “treatment” is the presence and action of a CEO and lead consultant at the Strategist action-logic or later and whose “effect” is organizational transformation.

The most significant threats to internal validity in such a study are the *interaction of selection and maturation*, *instrumentation*, *local history*, and *threats to statistical conclusion validity*. And the most significant threat to external validity comes in the form of *insufficient construct validity*, which is a subset of external validity (Cook & Campbell, 1979). Another important threat to the external validity of this study would appear to come from its *small sample size*.

The internal validity threat of *selection-maturation* would arise in this study if Strategist action-logic CEOs happened to be associated with types of organizations that had growth patterns systematically not encountered by the types of organizations headed by CEOs at earlier action-logics. In such a case, it could well be that extraneous causes, not CEOs’ and consultants’ action-logics, would account for the organizations’ transformation. In our study, however, there was considerable variety across the organizations that did and did not transform: a) in size (10-1,019 employees, average=485); b) in type (5 for-profit / 5 not-for-profit); and c) in line of business (investing, automobiles, energy, consulting, education, health care). In short, the successes and failures in organizational transformation are not associated with any of these variables (e.g., two of the three organizations that failed to transform were not-for-profits, but three of the five not-for-profits succeeded in transforming).

The threat of *instrumentation* arises when there are scaling problems with the measurement of the dependent variable (organizational transformation, in this case) such that changes are more likely to be measured in one group than the other. Looking into our organizations, we find differences in the baseline action-logics of organizational development of the different organizations, and we also find that the three organizations unsuccessful in transforming were among the four organizations in the study that began at a relatively late organizational action-logic. At first, this seems to suggest that the coding scheme the raters employed may not be sensitive to transformations beyond that action-logic, or that such late-action-logic transformation is much less likely to occur than transformations through the earlier action-logics. However, a closer look reveals that six of the seven organizations that were coded as having transformed actually progressed to the organizational action-logic beyond the one at which those three organizations began, thus showing that the dependent variable was in fact sensitive to such transformations and that they do occur with some frequency. In short, when organization transformation failed to occur as indicated by measurement methods, it is unlikely that a limitation in the measuring instrument was the cause.

Another credible threat to internal validity, *local history*, is troublesome if there are events exogenous to the study that affected only the experimental group and not the control group. Here, there were five experimental groups (the five organizations with Strategist CEOs) and five non-Strategist-led control-group organizations. As far as we can tell, this threat of *local history* is substantially eliminated by the variety in geography (multi-national), industry (six industries), and market niche of the ten organizations that cut across both experimental and control groups.

Lastly, *threats to statistical conclusion validity* may also endanger the internal validity of this study. Statistical conclusion validity concerns our ability to determine statistically significant (within a specified α level) co-variation between our independent and dependent variables (Cook & Campbell, 1979). In the focal study, the threats to statistical conclusion validity were minimized since the authors used the Spearman rank order test, which is the appropriate nonparametric statistical test, and found results that were statistically significant at the .01 level. (Note that nonparametric tests make fewer assumptions about the "normality" of the distribution and about uniform, interval distances between the numbers, and are, hence, less likely to make false assumptions.)

With regard to external validity, Rooke and Torbert's (1998) detailed discussion of how their operationalizations reflect their action-inquiry-informed theoretical constructs minimizes many of the threats to construct—and also, by definition, external—validity. In this study, the operationalization of the independent construct is in the form of the LDP in the period of the Cook-Greuter variant of the Washington University Sentence Completion Test (Cook-Greuter, 1999). The operationalization of the dependent construct is provided by the organizational development action-logics of the Developmental Action Inquiry theory (Torbert, 1976, 1987; Rooke & Torbert, 1998) and the reliability among the three raters.

But what about the *small sample size* in the 10-organization study? Isn't that a huge barrier to claiming that the results are in any way externally generalizable to other organizations? In fact, the answer to this question is "No." The small sample size did introduce a slightly higher risk of a Type II error (falsely rejecting a valid finding), since the statistical power is slightly less than the conventional .80. But this small-sample-size effect would have affected the interpretation of the results only if a significant correlation had *not* been found. Put differently, what an *n* of 10, accounting for 59% of the variance at the .01 level of statistical significance means is just the same as what an *n* of 1,000 at the .01 level of statistical significance means… namely, that the hypothesis is not disconfirmed, with less than one in a hundred chances that the inference is in fact false.

More than that, the small *n* of ten organizations in this study means that the psychometric measure of the CEOs' and the consultants' action-logic must be valid in virtually every single case and must be a very powerful causal variable indeed. Otherwise, neither the correlation, nor the test of significance would have been so strong. Put differently, if the *n* had been 1,000 organizations rather than 10 and the result had achieved the .01 level of statistical significance, then the hypothesis would be confirmed as it has been here, but the variable (CEOs' & consultants' action-logic, as measured by the LDP) would not have been demonstrated to be as powerful a causal factor as it has been demonstrated to be in this study.

At the same time, however, it is important to remain cautious about the generalizability of the findings in two regards. Since the largest business unit in the study had 1,019 employees, we cannot know whether the findings will hold for Fortune 500 size companies. (Indeed, we expect that the CEO's action-logic alone will account for less of the variance in huge organizations and that the action-logic of the top executive team will account for more.) Also, the organizations in this study, whether for-profit or not-for-profit, are all productive, economically-oriented, work organizations; hence, the findings may not be representative of all types of organizations (e.g., spiritual organizations, temporary political campaign organizations, families, or government agencies). On the other hand, the results should be generalizable to the 95% of business and competitive not-for-profit organizations that have 1,000 employees or less. (And… since most management research is conducted in Fortune 500 firms ([partly because they are the only ones with enough margin to be able to tolerate studies of absolutely-unknown-practical-value)], this study can potentially claim greater external validity than all of them, generalizable, as it statistically is, to 95% of all for-profit and not-for-profit organizations, rather than the usual 5%.)

Construct Validity of the Current Harthill LDP

Since 2005, Herdman-Barker, Rooke, and Torbert have further redefined the rating of the Alchemist action-logic to further align the LDP rating system with Developmental Action Inquiry theory. We have included not only Cook-Greuter's (1999) criteria for what she names the "Construct-Aware" action-logic, but also her criteria for what she names the "Unitive" action-logic, as necessary elements of the Alchemist action-logic. In addition, we have identified further criteria that we believe increase the probability of validly identifying responses generated by Alchemists' post-conceptual awareness, not just by lexical cognitive complexity. These changes are based on DAI theory (Torbert, 1991) that argues that the Alchemist action-logic is a post-structural position of continuing fluctuating awareness that easily recognizes all the action-logics from Impulsive to Elder in oneself and in social situations (for more detail, see Herdman-Barker & Torbert, 2010).

In 2008, Harthill sponsored a cluster analysis test, explicitly designed to test the construct validity test on the current Harthill LDP. The statistical technique of cluster analysis is designed to assess the extent to which a measurement tool is able to capture qualitative differences that are conceptualized in the underlying theory that the tool is trying to measure. To assess whether (and what kind of) a qualitative difference exists between Conventional and Post-conventional action-logics, according to the LDP, we analyzed the underlying pattern of two separate sub-samples: (a) 830 LDP protocols rated overall as “Conventional” (Achiever action-logic and earlier); and (b) 61 “Post-conventional” protocols ((Individualist or later). We found a striking difference between the patterns derived from these two sub-samples, illustrating the qualitative difference between Conventional and Post-conventional action-logics. (In other words, we are asking you, our readers, to attend with equal concern both to the quantitative element of this finding and to the qualitative, “presentational” [Reason, 1994] portrait of the findings below.)

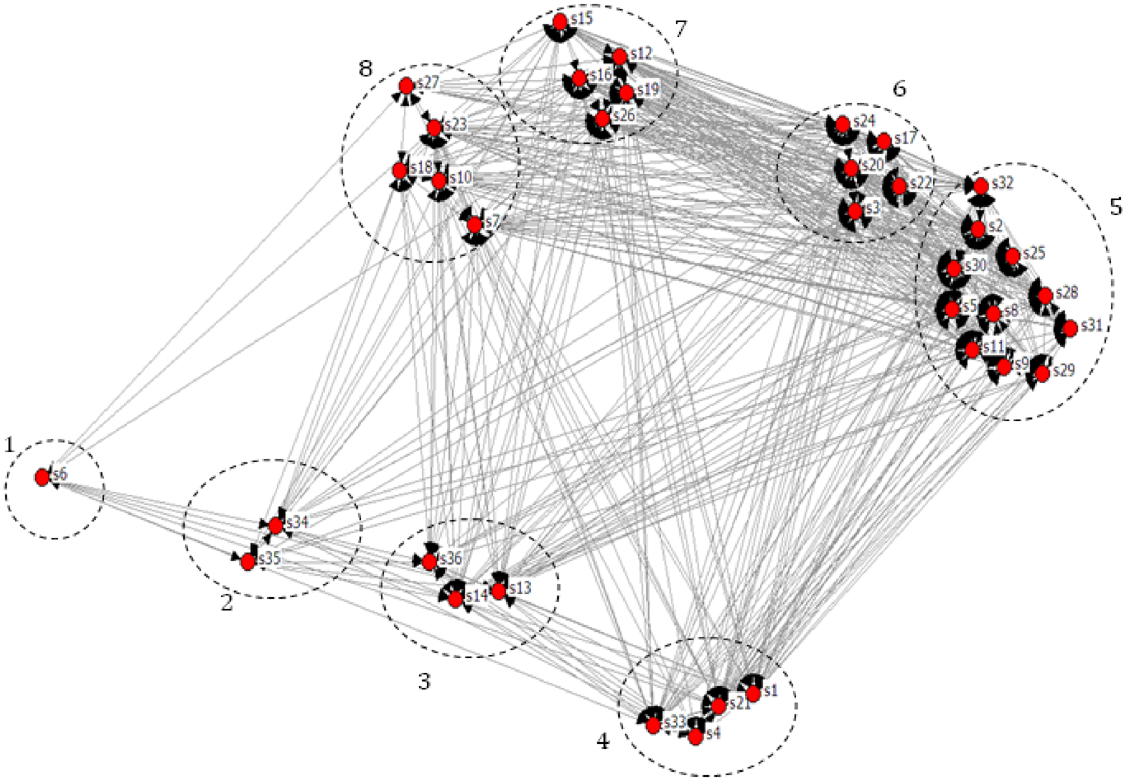

Figure 1. Cluster analysis of Harthill LDP scores on completions of 36 stems of Conventional profiles.

As Figure 1 shows, 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 LDPs is itself quite conventional statistically: distinct clusters or factors show up, with different sentence stems associated with each distinct factor.

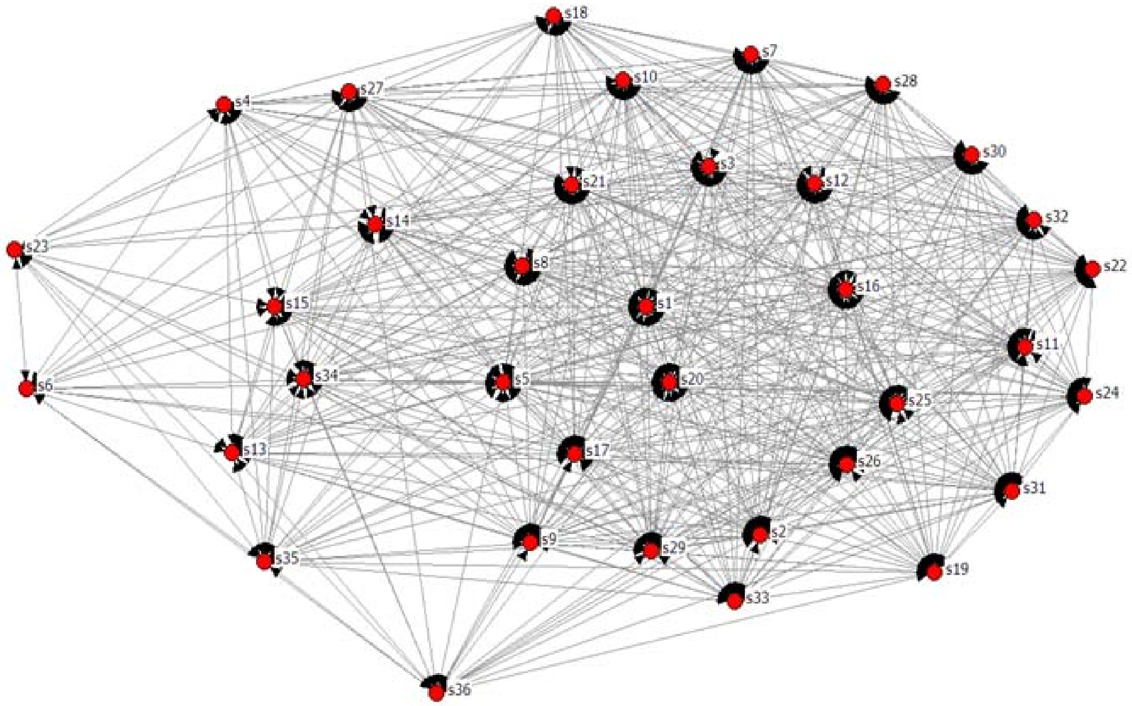

Figure 2: Cluster analysis of Harthill LDP scores on completions of 36 stems of Post-conventional profiles.

In contrast, we found a strikingly different pattern emerging from Post-conventional profiles (see Figure 2). For the Post-conventional action-logics (Individualist 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).

These results illustrate a fundamental difference between the Conventional and Post-conventional action-logics, echoing the adult developmental theoretical foundation on which the Harthill LDP is built (Kegan, 1982, 1994; Torbert, 1991; Torbert & Associates 2004). The stably-focused Conventional 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, "living" mental mapping process.

Plato's two distinctive images for the nature of thought in the *Theaetetus* – 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 difference between Conventional and Post-conventional thought.

Validity of the Harthill LDP on a Case by Case Basis

Because both the scientific and the pragmatic aim of using the LDP as one aspect of *Developmental Action Inquiry* theory, method and practice is to be able to support the further developmental transformations of those persons who take the measure and seek feedback about their performance on it, it is obviously crucial that the measure be as reliable and valid as possible on a case by case basis, and not merely on an aggregate statistical basis when comparing across large groups. The most recent reliability tests, cited at the end of the earlier reliability section, are performed on all late action-logic protocols partly in order to increase the likely validity in each individual case (and the associated very high reliability statistics also support the proposition that the LDP is accurate on a cases by case basis). Even more impressive are the very high correlations achieved in several of the validity tests cited above (Study 3 and the 10-organization study), which require case by case accuracy on the part of the LDP to generate such results. In addition, two recent doctoral dissertations have worked with small samples of late action-logic ratings and have shown that the measure accurately distinguishes among late action-logics on a case-by-case basis in theoretically predicted ways (McCallum, 2008; Nicolaides, 2008).

Post-Conventional Forms of Validity Testing

In addition to all of the foregoing conventional, third-person, social scientific reliability and validity tests of the LDP, Harthill also proposes, performs, and supports post-conventional, first-, and second-person validity testing of the measure each time it is offered as feedback to individuals using the measure. First, individuals are urged to make an estimate of their own center-of-gravity action-logic (e.g., based on reading Torbert & Associates, 2004) and to treat the subjective estimate as a form of first-person inquiry, triangling with the third-person, objective metric (Hartwell & Torbert, 1999a, 19989b). Furthermore, a 200-300-word personal commentary is provided with each profile, and the recipient is invited to read it in the following spirit: “This personalized commentary has been written in a spirit of invitation and inquiry (rather than as an authoritative statement of ‘the truth’). It is intended to provide support and provocation in exploring your thinking, actions and opportunities for growth.” Writing a developmental autobiography is an additional first-person method for exploring one’s transformation across center-of-gravity action-logics during one’s lifetime to date (Torbert & Fisher, 1992).

Second, individuals who have taken the LDP are urged to seek a 2nd-person form of inquiry and triangling as well. This can be an executive debriefing session with an authorized LDP debriefer, or an action-logic estimate based on the analysis, in a small community of inquiry, of the implicit framing assumptions a person makes in a difficult, work-or-personally-related conversation (McGuire, Palus & Torbert, 2007).

On the third-person scale of large organizations, institutions, economic sectors, or political regions, the aim of the DAI approach to theory, practice, and method is to build developmental research capable of generating single-, double-, and triple-loop feedback into the day-to-day operations of the institution. This enhanced feedback in the midst of everyone’s everyday life should be associated with increasing experiences of adult developmental transformation, with an increasing prevalence of organizational transformations to late action-logics, and with a transformation of social scientific inquiry itself toward later action-logics. Such widespread interweaving of first-, second-, and third-person development is a challenge for the future of social action and social inquiry.

Conclusion

Why is the Harthill LDP, based on the theory, practice, and method of Developmental Action Inquiry, so powerful in explaining real-life questions like which leaders succeed in supporting organizational transformation of organizations of moderate size?

We believe it is in part because of the empirical reliability and validity of the Harthill LDP as a metric, a well calibrated measure that is grounded in extensive empirical evidence of its reliability and validity, as gathered together in this article.

But the metric can only be so powerful because Developmental Action Inquiry theory describes distinctive action-logics through which persons and organizations predictably pass, as they become increasingly open and committed to integrating action and single-, double-, and triple-loop inquiry at the 1st-, 2nd-, and 3rd-person scales in the service of timely, mutually-transforming actions.

In general, what one sees in the transformation from the Loevinger WUSCT to the Harthill LDP, traced in this article, is typical of developmental transformations from Conventional action-logics to Post-conventional action-logics. First, the 3rd-person, Expert, Empirical Positivist scientific base (Torbert 2000a, b) of the original Loevinger WUSCT is preserved and enhanced over the years. Second, new, post-conventional action-logics are conceived, defined, and operationalized through Cook-Greuter's, Herdman-Barker's, Rooke's, and Torbert's work. Third, the 3rd-person measure is re-oriented so that it can play a role in a wider field where the effort is to integrate it with practitioners' 1st- and 2nd-person research and practices in the midst of daily work and life. Then it is tested in action for the rest of one's life, not least in the overlapping, de-centralized communities of inquiry one develops with friends to face the end of life together.

Thus, the Harthill LDP emerges from, and re-presents, a relatively late-action-logic Developmental Action Inquiry paradigm of social science and social action, wherein a psychometric measure is developed as-part-of-widespread-communities-of-inquiry-participating-in-an-integral-system-of-mutually-responsible-action-and-inquiry-in-the-present.

References

Chandler, D. & Torbert, W. (2003). Transforming inquiry and action: By interweaving 27 flavors of action research. *Journal of Action Research, 1,* 133-152.

Cohen, J. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences*. Hillsdale, NJ: L. Erlbaum Associates.

Cook, T. D. & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Boston: Houghton Mifflin.

Cook-Greuter, S. (1999). *Postautonomous ego development: A study of its nature and measurement.* Unpublished doctoral dissertation. Cambridge MA: Harvard Graduate School of Education.

Erikson, E. (1959). Identity and the life cycle. *Psychological Issues, 1,* 1-171.

Fisher, D. & Torbert, W. (1991). Transforming managerial practice: Beyond the achiever stage. In R. Woodman & W. Pasmore (Eds.), *Research in organization change and development*, *Vol. 5* (pp.143-173). Greenwich, CT: JAI Press.

Fisher, D. & Torbert, W. (1995). *Personal and organizational transformation: The true challenge of continual quality improvement*. London: McGraw-Hill.

Hartwell, J. & Torbert, W. (1999a). A group interview with Andy Wilson, founder and CEO of Boston Duck Tours, and Massachusetts Entrepreneur of the Year. *Journal of Management Inquiry, 8*(2), 183-190.

Hartwell, J. & Torbert, W. (1999b). Analysis of the group interview with Andy Wilson: An illustration of interweaving first-, second-, and third-person research/practice. *Journal of Management Inquiry, 8*(2), 191-204.

Herdman-Barker, E. & Torbert, W. (2010). Generating and measuring practical differences in leadership performance at postconventional action-logics: Developing the Harthill Leadership Development Profile. In A. Coombs, A. Pfaffenberger & P. Marko (Eds.), *The postconventional personality: Perspectives on higher development.* Albany NY: SUNY Academic Press.

Herdman-Barker, E., Livne-Tarandach, R., McCallum, D., Nicolaides, A. & Torbert, W. (2010). Developmental action inquiry: A distinct integral approach that integrates theory, practice, and research in action. In S. Esbjorn-Hargens, et al (Eds.), *Integral Theory in Action.* Albany NY: SUNY Press.

Kegan, R. (1982). *The evolving self.* Boston: Harvard University Press.

Kegan, R. (1994). *In over our heads: The mental demands of modern life*. Cambridge MA: Harvard University Press.

Kohlberg, L. (1963). The development of children's orientations towards moral order: I. Sequence in the development of moral thought, *Vita Humana*, *6,* 11-33.

Kohlberg, L. (1964). Development of moral character and moral ideology in M. Hoffman and L. Hoffman (Eds.), *Review of child development research*, *Vol. 1* (pp. 383-431). New York: Russell Sage.

Lorr, M. & Manning, T. (1978). Measurement of ego development by sentence completion and personality test. *Journal of Clinical Psychology, 34,* 354-360.

Livne-Tarandach, R. & Torbert, W. (2008). *One test of the validity of the Harthill Leadership Development Profile*. Unpublished manuscript. Available from torbert@bc.edu.

McCallum, D. (2008). *Exploring the implications of a hidden diversity in group relations conference training: A developmental perspective.* Unpublished doctoral dissertation. Columbia Teachers College, New York.

McCauley, C. Drath, W., Palus, C., O'Connor, P. & Baker, B. (2006). The use of constructive-developmental theory to advance the understanding of leadership. *The Leadership Quarterly*, *17,* 634-653.

McGuire, J., Palus, C. & Torbert, W. (2007). Toward interdependent organizing and researching. In A. Shani et al (Eds.), *Handbook of collaborative management research* (pp. 123 -142). Thousand Oak CA: Sage.

Merron, K. & Torbert, W. (1984). Offering managers feedback on Loevinger's ego development measure. Unpublished manuscript. School of Management, Boston College. Available from torbert@bc.edu.

Merron, K., Fisher, D., & Torbert, W. (1987). Meaning making and management action. *Group and Organizational Studies*,12, 274-286.

Molloy, E. (1978). *Toward a new paradigm for the study of the person at work: An empirical extension of Loevinger's theory of ego development.* Unpublished doctoral dissertation. University of Dublin, Dublin, Ireland.

Nicolaides, A. (2008). *Learning their way through ambiguity: Explorations of how nine developmentally mature adults make sense of ambiguity.* Unpublished doctoral dissertation. Columbia University Teachers College, New York.

Peterson, L. (2008). Clinical "significance:" "Clinical" Significance and "practical" significance are not the same things. Paper presented at the Annual Meeting of the Southwest Educational Research Association (New Orleans, LA, Feb 7, 2008).

Reason, P. (1994). *Participation in human inquiry*. London: Sage.

Reason, P. & Torbert, W. (2001). Toward a transformational social science: A further look at the scientific merits of action research. *Concepts and Transformation,6*(1), 1-37.

Rooke, D. & Torbert, W. (1998). Organizational transformation as a function of CEOs' developmental stage. *Organization Development Journal,16*(1), 11-28.

Schutt R. (2004). *Investigating the social world, the process and practice of research*, 4th ed. London: Pine Forge Press.

Rooke, D. & Torbert, W. (2005) Seven transformations of leadership. *Harvard Business Review,* April, 66-76.

Starr, A. & Torbert, W. (2005). Timely and transforming leadership inquiry and practice: Toward triple-loop awareness. *Integral Review 1*, 85-97.

Stein, Z., & Heikkinen, K. (2009). Models, metrics and measurement in developmental psychology. *Integral Review, 5*(1), 3-24.

Torbert, W. (1976). *Creating a community of inquiry: Conflict, collaboration, transformation*. London UK: Wiley Interscience.

Torbert, W. (1987). Education for organizational and community self-management. In S. Bruyn & J. Meehan (Eds.), *Beyond Market and State* (pp. 171-184). Philadelphia PA: Temple University Press.

Torbert, W. (1991). *The power of balance: Transforming self, society, and scientific inquiry.* Thousand Oaks CA: Sage.

Torbert, W. (1994). Cultivating post-formal development: higher stages and contrasting interventions. In M. Miller & S. Cook-Greuter (Eds.), *Transcendence and mature thought in adulthood* (pp. 181-203). Lanham MD: Rowman & Littlefield.

Torbert, W. (2000a). A developmental approach to social science: A model for analyzing Charles Alexander's scientific contributions. *Journal of Adult Development*, *7*, 255-267.

Torbert, W. (2000b). Transforming social science: Integrating quantitative, qualitative, and action research. In F. Sherman & W. Torbert (Eds.). *Transforming social inquiry, transforming social action* (pp. 67-92). Boston MA: Kluwer Academic Publishers.

Torbert, W. & D. Fisher (1992). Autobiographical awareness as a catalyst for managerial and organizational development. *Management Education and Development*, *23,* 184-198.

Torbert, B. & Associates (2004). *Action inquiry: The secret of timely and transforming leadership*. San Francisco CA: Berrett-Koehler.

Torbert, W, & Martinez, S. (in press). Toward a theory and practice of timely inquiry and action. *Journal of Action Research*.

Vaillant, G. & McCullough, L. (1987). The Washington University Sentence Completion Test compared with other measures of adult development. *American Journal of Psychiatry*, *144*(9), 1189-1194.

Webb, E., Campbell, D., Schwartz, R. & Sechrest, L. (1966). *Unobtrusive measures: Nonreactive research in the social sciences*. Chicago: Rand McNally.

Westenberg, M., Jonckheer, P., Treffers, P. & Drewes, M. (1998). *Personality development: Theoretical, empirical, and clinical investigations of Loevinger's conception of ego development.* Mahwah NJ: Lawrence Erlbaum.

Reut Livne-Tarandach, *Doctoral Candidate, Boston College: earned her MSc. in Organizational Studies from the Technion, Israeli institute of technology. She is currently a Ph.D. candidate in the Organization Studies department of the Carroll School of Management at Boston College. Her primary research interests are creativity, learning and organizational change. Her dissertation work is devoted to the intersection of organizational change and creativity where she explored how creative episodes are utilized by change agent groups to ignite, invigorate and revive change vitality. She has published in* Journal of Organizational Behavior, International Journal of Learning and Intellectual Capital, Research in Organizational Change and Development and Experimental Business Research: Marketing, Accounting and Cognitive Perspectives.

Bill Torbert *is a professor emeritus of leadership at Boston College who has authored many books and articles, including* The Power of Balance: Transforming Self, Society and Scientific Inquiry (1991) *and* Action Inquiry: The Secret of Timely, Transforming Leadership (2004). *Having consulted widely and served on boards of directors, he is currently Director of Research for Harthill Consulting UK.* torbert@bc.edu.