



University Learning Outcomes Assessment Validity

Introduction

The University Learning Outcomes Assessment (UniLOA) is specifically designed to measure college student behaviors along seven distinct yet inter-correlated domains including:

- Critical Thinking
- Self-Awareness
- Communication
- Diversity
- Citizenship
- Membership and Leadership
- Relationships

The UniLOA does not focus on academic-based, classroom content learning. To be sure, classroom-based learning, especially in the acquisition of disciplinary-specific learning is critical to overall student learning. Yet, that type of learning is in of itself insufficient to fully support holistic student development. Pascarella and Terenzini (1991) maintain that while classroom experiences have positive effects on gains, especially within disciplinary knowledge, those experiences alone have limited impact on learning from a holistic perspective. Rather, significant learning is the outcome of more generalized types of campus and community involvement and meaningful relationships with others; experiences that *are* measured by the UniLOA.

Reliability and Self-Report Instruments

Because the UniLOA relies on student self-reported behavior, it must be considered an indirect measure of student learning outcomes. The reliability of student self-report has received considerable attention in the past two decades as the use of assessment instruments relying on self-report has increased in both availability and use in higher education. The work of Dobbins, Jiing-Lih, and Werbel (1993) as well as Cassady (2001) reported that low-performing students tend to over report behaviors more than high-performing students, especially when they perceived specific levels of performance socially desirable. In his *attribution theory*, Heider (1958) as well as Weiner (1986) in his *attribution theory in education* proposed that individuals tend to overestimate their own performance accomplishment while underestimating the performance of others. Each of these theories would suggest that student self-report could be unreliable if it fails to equal true performance. However, many instruments, including the National Survey of Student Engagement (NSSE), the Community College Survey of Student Engagement (CCSE) College Senior Survey (CSS) College Student Experiences Questionnaire (CSEQ), among others, all employ self-report formats which have produced highly reliable data over time, suggesting that overestimation, if it does indeed occur, is fairly consistent across individuals and groups of students and therefore provides data against which accurate comparisons can be drawn, as Pike (1995) asserted. Additional information regarding the reliability of student self-reports can be found by referring to the thoughtful article by George Kuh, Shouping Hu, and Nick Vesper which refers specifically to the reliability of student self-report data in their 2000 Article, "*They shall be known by what they do:*" *An activities-bases typology of college students*" in the March/April, 2000 issue of the Journal of College Student Development.

The UniLOA as a Diagnostic and Prescriptive Tool

At present, the UniLOA serves both diagnostic and prescriptive functions. As a diagnostic tool, it presents mean scores to describe statistically normal behavior. Local scores can then be contrasted to national means, thus providing a “snapshot” of student behavior at the institutional or organizational level. As a prescriptive tool, the UniLOA provides scores for self-reported behaviors that institutions can analyze to answer two basic questions; 1) is the average level of behavior along the UniLOA’s 7 domains above, consistent with, or below articulated student learning outcomes, and 2) is there a service, support intervention or program (SSIP) that could (or should) be provided with the express purpose of bolstering student behaviors to a level consistent with a desirable level of articulated student learning outcomes? Specifically, the UniLOA identifies low-performance behaviors, thus providing critical information that can be used to guide institutional tactical and strategic planning.

The UniLOA as a Predictive Instrument

Because the UniLOA has been collecting data for only four years, not enough time has elapsed to support longitudinal studies to determine the degree to which the instrument can be used as a predictive tool in terms of college student persistence and/or graduation rates, and general life success beyond graduation. Until such time as reliable longitudinal studies can be completed, the UniLOA’s authors assert the validity of the assumption that behaviors resulting in positive outcomes tend to continue while behaviors eliciting negative consequences tend to decrease or be eliminated (Skinner, 1938).

The UniLOA Concept- Empirical Research

The UniLOA scales were constructed conceptually to determine primary domains of behavior considered critical to complement growth, learning, and development (GLD) of college students through graduation which prepares them for successful management of their lives professionally, socially, interpersonally, and intrapersonally. Identification of critical domains was made through two conduits, the first of which was an exhaustive review of contemporary research and literature in human development, specifically in the area of college student development, and through field studies designed to poll higher education constituents as to what they felt was important for holistic student GLD.

One of the most salient findings of the review of current research was the recurring theme of the high value of experiential learning (see Kendall 1990; Kendrick 1996; Warren 1998; Wink 2000) and the quality of GLD that occurs as a result of students’ engaging behaviors to better define, reinforce, and globalize classroom and out of classroom accomplishments. Consistent with empirical research, the UniLOA reveals difference in such demographics as gender (Watts and Bosshardt 1991, Baxter Magolda, M.B., 1992), age (Marlin and Niss, 1980, Seiver, 1983) and ethnicity (Fenwick, L.T., 1996, Peng, S. S., & Wright, D. (1994), among others.

The UniLOA Concept- Environmental Research

A second conduit employed to collect information resulted from a combination of electronically-based surveys, focus group meeting, and one-on-one structured interviews with higher education professionals, students, parents, employers, and other constituent groups of higher education by asking the basic question “what should a student possess in terms of skills, attributes, and qualities by the time they graduate?” Data collected from constituents was analyzed employing cluster analysis techniques to identify the most common observations of those polled. That analysis clearly identified the seven UniLOA domains. To be sure, other areas of concern were presented in the data, such as health and wellness, and commitment to lifetime learning, but no additional factors were revealed with the magnitude of the seven identified domains. After the initial identification of domains, each needed to be defined operationally. Those operational definitions ultimately became the basis for the development of the individual items’ construction.

The UniLOA Items

The UniLOA domains represent broad constructs which themselves cannot be measured. However, broad constructs can be operationalized by defining affective, cognitive, or behavioral phenomena associated with the construct. Each item of the UniLOA refers to a specific behavior positively correlated with the broader construct under which it falls.

Individual items on the UniLOA are carefully worded to present the underlying behavior as a cognitive root and a behavioral example provided as a stem. Use of a behavioral example assists test-takers in better understanding behavior positively correlated with the cognitive root.

The UniLOA's authors contend that while affective and cognitive states are critical to the lived experience, it is behavior that is observed by others and in the end, is the most important in producing outcomes, whether positive or negative.

The initial stage of item-development resulted in over 150 items being included in the piloted version of the UniLOA. The first stage of piloting was designed primarily for purposes of identifying and eliminating items that were unnecessary based on a number of criteria including;

- Very low standard deviations
- Significant single modal response sets at either end of the distribution
- Bimodal response patterns
- Redundancy

Item-reduction studies concluded with the elimination of many items, resulting in a set of 70 which now appear in the final version of the UniLOA.

Correlation Studies

In operationalizing the domains, it was clear that there would be considerable inter-item and inter-domain correlations. It was also noted that the pattern of inter-domain correlations would provide some face validity to the scales. For example, scores on the domains of Communication and Relationships ($r=.782$) should be more highly correlated than scores on Communication and Citizenship ($r=.563$)

UniLOA Inter-domain Correlation Table

	CT	SA	CS	D	CZP	ML	R	INTD
CT	1.000							
SA	.756	1.000						
CS	.803	.788	1.000					
D	.718	.709	.721	1.000				
CZP	.650	.610	.563	.639	1.000			
ML	.795	.780	.779	.746	.677	1.000		
R	.778	.799	.782	.766	.644	.847	1.000	
INTD	.824	.831	.923	.774	.546	.796	.817	1.000

Note – N=9946, all are significant at $<.001$

Because of the high degree of inter-item and inter-domain correlation, it was hypothesized that factor analysis would not reveal utilitarian results to describe the whole of student GLD. Yet, factor analysis with varimax rotation was conducted on the 70 retained items of the UniLOA. There was little expectation that any domain would emerge as an independent factor since the scales were so highly inter-correlated. Nine factors emerged with eigenvalues over 1.0 so a scree plot was used to assist in determining factors that should be retained, of which 2 emerged. The first factor had an eigenvalue of 23.3 and accounted for 33.3% of the variance. The second factor had an eigenvalue of 2.7 and accounted for an additional 3.9% of the variance.

The first factor became the eighth domain (Interdependence), with 15 items. Nine out of ten items on the Citizenship domain were those most highly correlated with the second factor, and the tenth item was among the 15 most highly correlated items with this factor. Because it introduces redundancy into the reporting of results, the Interdependence domain is not typically reported; however, it is computed and is available on request to UniLOA partners to assist in the analysis of results, should they have the desire.

Interpretation of UniLOA Data

Interpreting results of the UniLOA can be, like any research dealing with the complex phenomenon of the human experience, quite complex. While standard deviation is a preferred statistic to understand variance, it is most appropriate when responses are normally distributed. As raw scores of the UniLOA items and domains are converted, normal distribution would place the mean, median, and mode around a score of 55, but means for both domains and items on the UniLOA center around a score of 70 as a result of a large number of outliers; none of which are eliminated in computing mean scores as they are valid data points. The existence of outliers has a profound impact on the utility of using standard deviation to understand variance, rendering it ineffectual in the analysis of the UniLOA data set.

In as much as standard deviation is highly sensitive to outliers, it is not recommended for use in interpreting UniLOA results. While statistically simple, mean averages are best used to interpret results. Differences in mean scores of approximately 1.5 points begin to assume statistical significance and differences of 3 or more points assumes considerable practical significance.

Discriminant, Concurrent, Criterion and Predictive Validity

The strongest arguments for validity come from discriminant, predictive, and criterion-related validity. Questions on the UniLOA include 30 demographics items such as participants' gender and their average number of hours of sleep per night. Research on factors affecting student GLD was used to construct the list of 30 demographics questions, and analysis of the results indicate that different item and domain scores are impacted by different demographics.

Current content

Construct Validity

Construct validity refers to how well items measure or are correlated with the broader theoretical constructs, supported by theory and research. To assure high a high degree of construct validity, the UniLOA was developed by first examining available empirical research. Next, constituents having an interest in higher education were consulted through structured interviews, focus group meetings and surveys through comprehensive field studies.

Correlation Studies

Critical Thinking is most highly correlated with GPA (0.174), hours per week of study (0.144), accrued credit hours (0.128), gender (0.199), and having an academic scholarship (0.119). Negative correlations for gender and academic scholarship are an artifact of coding those variables as 0 and 1. Critical thinking scores are unrelated to average hours per night of sleep (0.009), number of organizational offices held (0.008), and hours per week spent watching TV or internet entertainment (0.058). On reflection, this supports concurrent and discriminant validity of the Critical Thinking domain on the UniLOA. It should be the case that critical thinking and academic achievement are related, and those students with more completed hours, or students who receive an academic scholarship should score higher. Consistent with empirical research on gender, the UniLOA reveals differences between men's and women's performance. ANOVA results show a significant yet small (1.72 points out of 70) difference between men's and women's performance. That the UniLOA is sensitive enough to discern differences between men's and women's performance is further argument for validity.

Self Awareness domain scores are most strongly associated with GPA ($r=0.190$), hours per week of study ($r=0.189$), the presence of an academic scholarship ($r=0.104$) and the number of organizational

memberships held in the last year ($r=0.090$). Age ($r=0.075$) and hours completed ($r=0.075$) are also correlated with Self Awareness demonstrating a slight association between the campus experience and Self Awareness. Not surprisingly, academically successful students engaging in academic pursuits score highly on self awareness. Scores on the Self Awareness domain are not associated with number of organizational offices held, military service, and hours per night of sleep providing, evidence for discriminant validity.

Communication domain scores are most highly correlated with GPA ($r=0.203$), hours per week of study ($r=0.132$), and having an academic scholarship ($r=0.130$), furthering the argument that the underlying construct of the Communication scale is similar to the underlying personal behaviors, attitudes, and beliefs that lead to superior academic performance. Scores on the Communication domain are least correlated with number of organizational offices held ($r=0.024$), hours per week of paid work ($r=0.025$), gender ($r=0.036$), and hours per week watching TV or on-line entertainment ($r=0.038$) demonstrating that the underlying construct is not related to these activities.

Diversity domain scores are most highly correlated, like the previous domains, with hours per week of study ($r=0.112$) and GPA ($r=0.102$), and is also correlated with hours per week participating in volunteer activities ($r=0.107$) and number of organizational memberships ($r=0.094$). The presence of volunteer activity and organization memberships, along with the expected academic performance correlations, indicates the importance of the relationship between volunteer activity and organizational membership experiences. Diversity scale scores also tap into the same underlying academic performance construct as the previous domains. Scores on diversity are least correlated with number of organizational offices held ($r=0.016$), military service, ($r=0.024$) and gender ($r=0.044$).

Citizenship domain scores are most correlated in a different pattern than the other scales, perhaps because of the different factor loading for the items on this domain. Scores on the Citizenship domain are most highly correlated with number of organizational memberships ($r=0.187$), hours per week of study ($r=0.140$), hours per week of volunteer activity ($r=0.131$), hours completed ($r=0.115$) and number of organizational offices held ($r=0.112$). This would indicate a relationship between the Citizenship domain scores and other behaviors associated with formal membership in the communities in which students are members. Scores on the Diversity domain are most weakly associated with living on or off campus ($r=0.016$), hours per night of sleep ($r=0.006$), number of roommates ($r=0.007$), military status ($r=0.007$) and Pell Grant status ($r=0.008$).

Membership and Leadership domain scores are most highly correlated with the familiar array of academic performance behaviors like GPA ($r=0.156$), number of organizational memberships ($r=0.153$), and hours per week of study ($r=0.146$). Hours completed ($r=0.115$) is an indicator that students with more hours score more highly on Membership and Leadership, giving argument to the value of the college experience, and the correlation with hours per week of volunteer activity ($r=0.106$) further strengthen the argument that some college experiences have a positive impact on student learning outcomes. Scores on Membership and Leadership are least associated with academic major ($r=0.032$), hours per night of sleep ($r=0.023$), and hours per week of paid work ($r=0.037$).

Relationship scale scores are most highly associated with hours per week of study ($r=0.144$), GPA ($r=0.140$), and number of organizational memberships ($r=0.105$), again reflecting the array of academic work and campus involvement behaviors reflected in the other scales. Relationship scores are least associated with academic major ($r=0.014$), hours per night of sleep ($r=0.015$), and number of organizational offices held ($r=0.024$).

Concurrent Validity

Concurrent validity refers to how consistent a particular instrument's results are when contrasted with other instruments designed to measure similar constructs. The establishment of concurrent validity is an on-going process and the UniLOA has already produced data and interpretive results that are highly consistent with the findings of other research and instruments designed to explore the student experience

in higher education settings. Please see comments under the previous section regarding previous empirical research.

Criterion-Related and Predictive Validity

Criterion-related validity produces the most utility in that it presents the correlation between test scores and actual outcomes over time. In the UniLOA, positive correlations between such variables as retention, persistence, and ultimate graduation, along with other variables that are less “clear” but none the less important such as post-graduate entry into a career field within the academic major, effectiveness in job performance, contribution to the community, and so on are the ultimate measures of criterion-related validity. Unfortunately, UniLOA data has not been collected for a long enough period of time to engage in meaningful longitudinal studies. Establishing criterion-related validity is an ongoing process which will become more precise over time.

The consistent association, though small in effect size, between academic behaviors like hours studying and GPA as a measure of academic success and scores on the seven UniLOA domains provides support for validity in that the domains are associated with measures of academic success, but only to a slight degree. Further, the association between domain scores and participation in campus organizations provides evidence of validity that the UniLOA is sensitive to co-curricular activities as well as academic activities.

References

- Baxter Magolda, M.B. (1992). *Knowing and reasoning in college*. Jossey-Bass: San Francisco.
- Cassady, J. C. 2001. Self-reported GPA and SAT: A methodological note. *Practical Assessment, Research & Evaluation* 7(12).
- Dobbins, Gregory H., Jiing-Lih Farh, and James D. Werbel. 1993. The influence of self-monitoring and inflation of grade-point averages for research and selection purposes.” *Journal of Applied Social Psychology*. 321-334.
- Fenwick, L.T. (1996). *A perspective on race equity and science and math education: Toward making science and math for all*. Paper presented at the Annual Conference of the Georgia Initiative in Mathematics and Science, Atlanta, GA.
- Heider, Fritz. (1958). *The psychology of interpersonal relations*. New York: John Wiley & Sons.
- Kendall, Jane C. (Ed.) (1990). Combining service and learning: An introduction. In *Combining service and learning: A resource book for community and public service*, 1-36. Raleigh, NC: National Society for Internships and Experiential Education.
- Kendrick, J. Richard. 1996. Outcomes of service learning in an introduction to sociology course. *Michigan Journal of Community Service Learning*, (3), 72-81.
- Marlin, J. W. and Niss, J. F. (1980). End of course evaluations as indicators of student learning and instructor effectiveness.” *Journal of Economic Education* (Spring): 16-27.
- Pascarella, E. T., & Terenzini, P. T. (1991). *How college affects students: Findings and insights from 20 years of research*. San Francisco: Jossey-Bass.
- Peng, S. S., & Wright, D. (1994). Explanation of academic achievement of asian american students. *Journal of Educational Research*, 87 (6), 346-352.
- Pike, G. R. (1995). The relationships between self-reports of college experiences and achievement test scores. *Research in Higher Education*, 36.

Seiver, Daniel. 1983. "Evaluations and grades: A simultaneous framework." *Journal of Economic Education*. (Summer). 32-38.

Skinner, B.F. (1938). *The behavior of organisms*. New York: Appleton-Century-Crofts.

Warren, Karen. 1998. Educating students for social justice in service learning. *Journal of Experiential Education*. 21 (3): 134-39.

Watts, M. & Bosshardt, W. (1991). "How instructors make a difference: Panel data estimates from principles of economics courses." *Review of Economics and Statistics* (May): 336-340.

Weiner, B. (1986). *An attributional theory of motivation and emotion*. New York: Springer-Verlag.

Wink, Joan. 2000. *Critical pedagogy: Notes from the real world*. (2nd ed.). New York: Addison-Wesley Longman.

