COOPERATIVE INSTITUTIONAL RESEARCH PROGRAM
at the HIGHER EDUCATION RESEARCH INSTITUTE AT UCLA

CIRP Construct Technical Report

Jessica Sharkness
Linda DeAngelo
John Pryor

Higher Education Research Institute
Graduate School of Education & Information Studies
University of California, Los Angeles

January 2010

Table of Contents

Introduction

We use questionnaires to gather information about phenomena of interest. In educational research we might be asking questions about how students interact with faculty, or how satisfied they are with college, or what kind of values they hold. In many cases this is how we who do survey design start thinking about what we want to know, with broad questions that embrace concepts that are multifaceted. But, we cannot ask students "how do you interact with faculty," because the question is too vague and also too broad, and we may not find out the full range of interactions students have with faculty. Thus, in order to learn about student and faculty interaction, and provide more context to students who are completing our questionnaires, we ask questions about different types of interactions that are important (e.g., conducting research with faculty), skipping those that are less important (e.g., passing a faculty member in the hallway without saying anything). This process provides us with a bank of items that cover what we believe are the important aspects of student-faculty interaction; taken together, the items can tell us about the state of student and faculty interaction generally.

There are several points that are important here. One is that the reason that we ask all of these detailed questions is not only to gather information about specific behavior, but also to get at the more elusive concept underlying the questions, often referred to as a latent trait. In combination, a set of items can provide a fuller understanding about an underlying latent trait than can any item individually. There are many different ways to combine survey items, and while all methods of data reduction are intended to help organize the information from survey items into smaller more useful chunks, just exactly which method we use is important. Each data reduction method has different implications in terms of how valuable the final combined piece of information is for its intended purposes.

Researchers using CIRP data have for decades used data reduction techniques to make sense of survey data, most often using techniques that fall under the rubric of Classical Test Theory (CTT). Much of the knowledge of the factors (what we term constructs) that these researchers developed, however, has stayed in professional journals, and has seldom made it into practice at the institutional level. Furthermore, most constructs were created individually by researchers on an ad-hoc basis, and there was no standardization across surveys or data sets in terms of which items were included in which constructs. We felt that colleges and universities that have CIRP data would benefit from the creation of a set of standard measures that are constant across survey instruments and across survey years. Not only would these measures help institutions to assess important latent traits among their students, but they would also help to reduce and organize the available information contained in each of our surveys.

We at CIRP embarked on a project, therefore, to organize and evaluate all of the latent traits that have been assessed using CIRP data, and to create a set of statistically sound, educationally relevant constructs to be used by institutions and researchers alike. The first part of this project was an exhaustive literature review of all of the research studies that have used CIRP data to generate measures of latent traits (constructs). While many constructs had been created over the years, covering many topical areas, we discovered that there were no universal "core" set of measures that were available to use. Each researcher created their own constructs specifically crafted to the dataset and population being studied. The second part of the project was an investigation into what are the best, most modern statistical methods for combining items into measures of the underlying latent traits of interest in our surveys. The result of this investigation was a decision to use Item Response Theory (IRT) rather than Classical Test Theory (CTT) to develop the constructs. Once these two questions were answered, we went

about creating a set of well-defined and well-measured constructs that are to be provided in each CIRP database. These constructs are designed to be used both locally, at an institution, for internal assessment, as well as more broadly, by researchers using the aggregate national data. Thus the goal of this project was to end up with a set of CIRP Constructs in each of the CIRP survey databases that could help guide research and our understanding of the college experience. This technical paper describes the second part of the process that we went though in creating the CIRP Constructs.

This report begins with a discussion of Classical Test Theory (CTT) and Item Response Theory (IRT), and then reviews the methods we used to create the CIRP Constructs. This review uses information on how we build the *Student-Faculty Interaction* Construct on the Your First College Year Survey (YFCY) as an IRT methods example. This report concludes with an appendix which includes detailed information about each of the CIRP Constructs, including construct definitions, survey items, and scoring parameters. In addition, we offer on our website answers to frequently asked questions about the Constructs that are beyond the scope of this technical report.

Classical Test Theory and Item Response Theory

Classical Test Theory (CTT) and Item Response Theory (IRT) are the two primary measurement theories that researchers employ to construct measures of latent traits. Because latent traits are by their very nature unobservable, researchers must measure them indirectly through a test, task, or survey. The reason unobservable traits can be accessed in such a way is that the traits are assumed to influence the way that people respond to test or survey questions. A variety of items are needed to measure a single latent trait, because one individual question cannot tell a researcher much more than what was asked. For example, if a survey item asks a student how often he or she "analyzed the basic elements of an idea, experience, or theory," then the response to that item would tell the researcher just that: how often a student believes he or she analyzed the basic elements of an idea. If what a researcher really wants to measure is the "higher-order thinking activities" of a student, however, then the researcher would need to ask additional questions, such as how often the student "made judgments about the value of information, arguments or methods," "applied theories or concepts to practical problems in new situations," and "synthesized and organized ideas, information, or experiences into new, more complex interpretations and relationships." The researcher could combine then the responses to all these questions into a scale representing the larger construct (c.f. Pascarella, Cruce, Umbach, Wolniak, Kuh, Carini, Hayek, Gonyea & Zhao, 2006).

No perfect measure of a latent variable can ever exist. By examining how a person responds to a set of items relating to a single underlying dimension, however, researchers can create scores that approximate a person's "level" of the latent trait. CTT and IRT are both tools that can be used to do this, but beyond their similar purpose the two measurement systems are quite dissimilar. CTT and IRT differ significantly in their modeling processes and they make

fundamentally different assumptions about the nature of the construct being measured as well as about how individuals respond to test items. A more in-depth treatment of CTT can be found in Lord and Novack (1968) or Allen and Yen (1979/2002), and more detail about IRT can be found in Embretson and Reise (2000). Below, a very basic outline of each theory is sketched in order to compare the two as they relate to the measurement of constructs covering the college experience.

Perhaps the most fundamental assumption of CTT is that a respondent's observed score on a scale or test represents his or her "true" score plus random error. The true score is a theoretical construct defined as "the mean of the theoretical distribution of…scores that would be found in repeated independent testings of the same person with the same test" (Allen & Yen, 1979/2002, p. 57). Error consists of random, unsystematic deviations from true score that occur in each testing occasion. Because error is random, it varies in every test administration, and as a consequence, observed score does also. True score, by contrast, is theoretically the same regardless of testing occasion. This does not mean, however, that a person's true score is "true" for every test or measure of the same construct—it is simply "true" for that person taking one specific test. That is, true scores are tied to a specific set of items as opposed to a "real" latent trait. If two tests of math ability have different questions and/or a different number of items, and Joe, who has some constant latent level of math ability, took both tests, he would have a different "true" score for each test because of the different forms. CTT estimates of traits, then, are test-dependent, and every test or scale has different psychometric properties.

The fundamental assumption underlying IRT is that every respondent has some "true" location on a continuous latent dimension (often called "theta," or θ). This location theta is assumed to probabilistically influence a person's responses to any item or set of items on a survey or test that covers the trait that theta represents. IRT models theta by using mathematical

equations that relate response patterns to a set of items, the psychometric properties of these items, and knowledge of how item properties influence responses (for more details see Embretson & Reise, 2000). Embretson and Reise (2000) describe IRT as being "akin to clinical inference" (p. 54); IRT provides a 'diagnosis' (trait estimate) for a person based on observed 'symptoms' (response patterns) and background knowledge (a mathematical model). There are a variety of different IRT models that can be used to explain how items influence response behavior and how best to estimate theta; the choice of these depends on the nature of the data to be analyzed.

There are several differences between CTT and IRT that are important for researchers measuring the impact of the college experience using scales from student surveys. First, in CTT a person's "true score" is entirely dependent on a particular set of items because the true score is defined in relation to a specific test or scale. In IRT, a person's "true score" is entirely independent of items because the underlying dimension of interest is only assumed to influence—it is not defined by—responses to specific items. Second, IRT explicitly models the relationship between person properties and item properties with the same model, while CTT does not. This means that score interpretation in IRT can be more interesting and flexible. For example, specific item responses can be directly compared to student's trait estimates in IRT, so what it means to be "high" in involvement, for instance, can be defined by specific activities. In CTT, scores can only be compared to other scores, so to interpret a score (is a score of 50 high or low?) reference must be made to a norm group (i.e. how many people scored above 50?). Third, the standard error of measurement (SEM) is treated differently in CTT and IRT. Because of assumptions made about measurement error in CTT (i.e. that it is normally distributed within persons and homogeneously distributed across persons), a test or scale's reliability and SEM are

estimated as a constant for all test-takers (Allen & Yen, 1979/2002). IRT, by contrast, allows for the possibility of different scale SEMs for different values of theta, and allows items to differentially affect SEM depending on how they relate to theta. The latter is a more flexible approach and likely more realistically approximates how people respond to tests and surveys. It also allows researchers to construct scales that maximally differentiate people from one another, either across the entire theta continuum or on some critical area of the continuum. Finally, a consequence of all of the above is that CTT scale scores and their interpretation are always context specific; in particular, they are item- and sample-specific. In IRT, the reverse is the case: item parameters are independent of sample characteristics, and theta estimates are independent of specific items. Given the selection of an appropriate IRT model, responses from any set of relevant (calibrated) items can be used to estimate a person's theta.

Methods

Each CIRP construct was created using the same general process. Below we describe the steps taken to develop each construct, and then illustrate the process using the *Student-Faculty Interaction* Construct from the YFCY.

Step 1: Item Selection and Assumption Checking

Initial Item Pool. Before any statistical analyses could be run, a pool of survey items that covered the relevant area of interest had to be identified. The selection of initial item pools for all of our constructs was guided by previous work from CIRP researchers as well as Astin's involvement theory (1984/1999), which defines college student involvement as "the investment of physical and psychological energy in various objects" on campus, which "may be highly generalized (the student experience) or highly specific (preparing for a chemistry examination)" (p. 519).

Exploratory factor analyses for item selection and assumption checking. The items in the initial pools were next evaluated via exploratory factor analysis to determine each item's fitness as an indicator of the construct of interest. The goal of factor analysis in this context is to determine whether the variance shared by a set of items can be explained by a reduced number of latent variables (factors). Specifically, in evaluating our initial item pools we were interested whether the interrelationships between the proposed variables in each scale could be best explained by one and only one underlying factor (Clark & Watson, 1995; Cortina, 1993; Gardner, 1995; Reise, Waller & Comrey, 2000; Russell, 2002). Note that due to the ordinal nature of item responses on the CIRP surveys, we used polychoric correlations for all relevant analyses in the place of the more traditional but less appropriate Pearson correlations (for more information see Dolan, 1994; Jöreskog & Sorbom, 1989; Olsson, 1979). All polychoric

correlations were computed using the software R 2.9.0 (R Development Core Team, 2009) and the maximum likelihood estimation algorithm in the polycor package (Fox, 2009). R was also used, along with Reville's psych library (2009), to conduct exploratory factor analyses. Following Russell (2002)'s recommendations, these exploratory analyses employed principal axis factoring with promax rotation (an oblique rotation).

The exploratory factor analyses performed for item selection was one of the most critical steps of the construct development process. Not only did the analyses result in the final selection of items for each construct, but they also constituted checks for some of the most fundamental assumptions of IRT. Two major assumptions underlie the estimation of appropriate item parameters in IRT: (1) local independence and (2) unidimensionality (actually, the assumption is "appropriate" dimensionality; here we are interested in only one dimension so we focus on unidimensionality; for more details see Embretson & Reise, 2000). The assumption of local independence dictates that the interrelationships among items in a scale be due only to the fact that they tap into the same underlying trait of interest. That is, local independence will be obtained if responses to the items in a scale are unrelated (independent of one another) once the underlying trait is controlled for. The assumption of unidimensionality is closely related to local independence, and in fact it will be satisfied if the local independence assumption is satisfied based on a single factor solution. Unidimesionality means that a single latent trait underlies the probability of responses to all items in a scale. When unidimensionality is met, score estimates will be "unambiguous indicators of a single construct" (Embretson & Reise, 2000, p. 227); when it is violated scores will reflect the influence of two or more dimensions.

During the process of performing the exploratory factor analyses we took several indicators into consideration to ensure local independence and unidimensionality. Specifically,

for each scale we: a) examined several different factor solutions (one factor, two, three, etc.) to ensure that the one-factor solution was most appropriate for the collection of items; b) created and inspected a Scree plot of the eigenvalues for each group of items to visually confirm that the data dictated a one-factor solution (Cattell, 1966); and c) compared a model-reproduced correlation matrix based on a one-factor solution to the observed correlation matrix to ensure that the resulting residual correlation matrix was composed of residuals that are small (< .10) and clustered around zero. If the differences between the single-factor model-reproduced correlations and the observed correlations are small and are clustered closely around zero, it can be said that the single factor solution is appropriate (McDonald, 1982; Reise, Waller & Comrey, 2000; Tabachnick & Fidell, 2007). The process of examining these indicators was iterative, for we were interested in reducing the initial hypothesized set of items to a group that met all of the conditions specified above. We first performed the exploratory factor analyses on every item in the initial item pool for each construct. If anomalies were found, that is, if a one-factor solution was not the most appropriate for the entire set of items, single items were removed one by one until a satisfactory solution could be obtained. An example of how this was done can be found in the "Example" section below.

Step 2: Parameter Estimation

Graded Response Model. Because the items in the all of CIRP's constructs are coded into ordinal categories, scored on Likert scales, the appropriate IRT model to use is Samejima's (1969) graded response model (GRM) (Embretson & Reise, 2000; Ostini & Nering, 2006). We applied the GRM model to our data using *MULTILOG 7* (Thissen, Chen, & Bock, 2002). The process of estimating parameters in the GRM is too complex to describe here, but excellent treatments can be found in Embretson & Reise (2000) and Ostini & Nering (2006). What is

important to note is the two types of parameters that result from applying the GRM. First, each item (i) has a discrimination or "slope" parameter, represented by α_i, which provides an indicator of how well an item taps into construct of interest. Items that have higher discriminations (α's) provide more information about the trait; in many respects these parameters are similar to factor loadings or item-total correlations. Discrimination parameters above 1.70 are considered very high, those between 1.35 and 1.70 are high, and those between .65 and 1.34 are moderate (Baker, 2001).[1]

Each item also has a series of threshold parameters associated with it. The number of threshold parameters for an item is equal to the number of item response categories minus one (k-1); the thresholds are here represented as $\beta_{i,1}$, $\beta_{i,2}$…$\beta_{i,k-1}$. The threshold parameters (β's) are given on the same metric as the underlying trait (θ), which for model identification purposes is assumed to have a standard normal distribution with a mean of 0 and a standard deviation of 1 (Embretson & Reise, 2000). Threshold parameters can be interpreted as the points on the latent trait continuum (e.g. the "level" of the trait) at which a respondent has a 50% probability of responding to an item in a certain response category or above and a 50% of responding in any other lower category (Embretson & Reise, 2000). For example, if a three-category item $_i$, such as one that has response options of never, occasionally and frequently, has a $\beta_{i,1}$ of -2.0 and a $\beta_{i,2}$ of 0.0, this means that the model predicts that a respondent with a level of the relevant latent trait two standard deviations below the mean (θ = -2.0) has a 50% chance of responding in the first category (never) and a 50% chance of responding in the second or third category

[1] Note that these numbers assume that the α's were estimated using a logistic function that does not include a D = 1.7 constant in the numerator of the equation. The inclusion or exclusion of this constant is unimportant in terms of the discussion in this paper, as it has to do with equating normal ogive functions and logistic functions and does not affect the parameter estimation procedure. However, it does affect parameter interpretation. Specifically, when a model that estimates item parameters does not include the D= 1.7 constant, the α's that are estimated are higher by a magnitude of 1.7 as compared to those estimated by a model that includes the constant. See Embretson & Reise, 2000 and Ostini & Nering, 2006 for more details.

(occasionally/frequently), while a respondent with a latent trait level at the mean ($\theta = 0.0$) has a 50% chance of responding in the first or second category (never/occasionally) and a 50% chance of responding in the third category (frequently). Respondents who fall below -2.0 on the latent trait level are most likely to respond "never," those between -2.0 and 0.0 are most likely to respond "occasionally," and those above 0.0 are most likely to respond "frequently." The amount of information an item provides about any given area of the latent trait depend on the value of the $\beta_{i,k-1}$'s and on how clustered or spread out they are.

Reference Population used for parameter estimation. All of the parameters were estimated using actual CIRP data from either 2008 or 2009. Data from 2008 were used for almost all constructs except for those that contained items not included on the surveys in 2008; in this situation 2009 data were used instead. When a construct was being developed only for one survey (e.g., just the TFS, YFCY or CSS), we used all the data from that year to perform parameter estimation. When we planned to create a construct that spanned more than one survey instrument, we created a dataset composed of equal numbers of students from each of the relevant survey databases. Typically this involved combining all YFCY and/or CSS cases with a random sample of TFS cases equal to the number of YFCY/CSS cases. Estimating the parameters in this way ensured that constructs would be able to be compared on the same metric across survey instruments and populations.

Final Parameters. The final parameters for all of the CIRP Construct items are listed in the Appendix of this report.

Step 3: Scoring

MULTILOG Scoring. Using the parameters estimated for each construct, MULTILOG was again used to score students on each construct. Scoring in IRT is an iterative process that

finds the most likely trait level for a student, given the responses he or she gave to the set of questions that define a construct. Embretson and Reise (2000) explain score estimates as follows: "for every position on the latent-trait continuum, from positive to negative infinity, a likelihood value can be computed for a particular item response pattern…another way of phrasing this is to ask, given the examinee's pattern of…responses to a set of items, with assumed known item parameter values, what is the examinee's most likely position on the latent-trait continuum?" (p. 159). Due to problems with estimating scores for respondents who respond all in the highest or lowest categories, MULTILOG incorporates a prior distribution for the latent trait into the score estimating process (the process is called Maximum A Posteriori Scoring, or MAP). In IRT the metric of this prior latent distribution is arbitrary; MULTILOG sets it as a standard normal, with a mean of 0 and a standard deviation of 1 (Thissen, Chen, & Bock, 2002). Therefore the scores assigned to each response pattern/respondent are also given on this distribution.

Rescaled Scores. Students' scores were initially given on a "z-score" metric. Although statisticians often work in standardized z scores, these scores are not always the most ideal for interpretative purposes given the decimals in such scores as well as the negative scores for half the population. Therefore, before merging students' scores with their CIRP data, we rescaled all students' scores to be on a mean of approximately 50 with a standard deviation of approximately 10. This was done by multiplying each score by 10 and adding 50. These are the final scores that are appended to each CIRP data set and that are provided in our newly revamped reports.

Score Categories. The CIRP reports also describe the constructs using a three-category variable, labeled "low," "medium," and "high." This variable is created by recoding the original, continuous scores according to their observed distributions (means and standard deviations). Students with scores of 0.5 standard deviations above the mean or higher are coded into the

"high" category; students with scores within 0.5 standard deviations of the mean are coded into the "medium" category; and students with scores of 0.5 standard deviations below the mean or lower are coded into the "low" category.

Example of the construct creation and scoring process: First-Year Faculty-Student Interaction (YFCY)

Step 1: Item Selection and Assumption Checking

Initial Item Pool. In the example here, first-year faculty involvement was conceptualized as a combination of the quantity and quality of faculty-student interaction. The construct specifically measures the amount and type of faculty contact that students have during their first year of college, as well as satisfaction with these issues. Table 1 below lists all of the items from

Table 1
All 2008 YFCY Items Relating to Faculty Involvement

Scale/Item	Response Options
Since en tering this co lle ge, how often have you interacted [by phone, e-mail, Instant Messenger, or in person] w ith …Faculty outside of class or office hours	Daily, 2 or 3 times per week, once a week, 1 or 2 times per month, 1 or 2 times per term, Never
Since en tering this co lle ge, how often have you interacted [by phone, e-mail, Instant Messenger, or in person] w ith …Faculty during office hours	
Since entering this co llege, how often have you …Asked a professor for advice after class	Frequently, Occasionally, Not at all
Since en tering this co lle ge, how often have you received fro m your professor …advice or guidance about your educational program	
Since en tering this co lle ge, how often have you received fro m your professor …emotional support or encouragement	
Since entering this college, have you… communicated regularly with your professors	Yes (2) , No (0)
Please ra te your satisfa c tion with th is ins titution [in terms of the]… Amount of contact with faculty	Very Satisfied (5), Satisfied (4), Neutral (3), Dissatisfied (2), Very Dissatisfied (1), Can't Rate/ No Experience (missing)
Since en tering this co lle ge , how much time have you spent during a typica l we ek … talking with professors outside of class	None, Less than 1 hour, 1-2, 3-5, 6-10, 11-15, 16-20, Over 20

the 2008 YFCY that were related to faculty involvement.

Exploratory factor analyses for item selection. Initial exploratory factor analyses were run on the full set of faculty involvement variables listed above. Based on the results of these analyses, three items were removed from the faculty involvement item pool. The first item to be removed from the item pool was the question asking about the number of hours per week students typically spent talking with faculty outside of class. This was removed because it was deemed to be essentially the same question as several of the others—specifically, it conceptually overlapped too much with frequency of interacting with faculty during office hours, frequency of interaction outside of class or office hours, frequency of asking a professor for advice after class, and communicating regularly with professors.

Next, the variable representing the frequency with which students received from professors emotional support or encouragement was removed. The reason this variable was dropped was due to what is called a "local dependence"—a violation of one of the assumptions of IRT. Specifically, the "professors provide emotional support or encouragement" variable had an extremely high correlation with the "professors provide advice or guidance about educational program" variable ($r = .65$), and this correlation was unexplained by a factor model assuming only one underlying latent trait (unexplained (residual) correlation based on a one factor solution = .22). Therefore, one of the variables had to be removed to avoid violating the local independence assumption of IRT. We kept the advice about educational program variable instead of the emotional support variable because we deemed the former type of faculty support more directly related to the types of interaction we expected between students and faculty in the first year of college.

Finally, the variable asking about the frequency with which students interacted with

faculty during office hours was recoded into a dichotomous variable representing whether students ever went to office hours. The variable was coded this way because we desired to keep a measure of going to office hours in the overall construct, but the full variable could not be included due to, again, a local independence violation between it and the variable "How often in the past year did you interact with faculty outside of office hours?" (The correlation between the original two variables was .55 and the unexplained correlation based on a one-factor solution was .22; the correlation between the "office hours yes/no" variable and the interaction with faculty outside of office hours variable was .38 and the unexplained correlation was .07). Table 2 below shows the final items in the faculty involvement scale and the polychoric correlations between them.

Table 2
Polychoric correlations between final faculty involvement items

		1	2	3	4	5	6
1	Freq: Interact with faculty outside class/office hours	1.00					
2	Frequency: Asked a professor for advice after class	0.39	1.00				
3	Yes/No: Communicate regularly with your professors	0.46	0.53	1.00			
4	Satisfaction: Amount of contact with faculty	0.31	0.33	0.51	1.00		
5	Freq: Prof. provide advice about educational program	0.33	0.47	0.51	0.41	1.00	
6	Yes/No: Ever go to office hours	0.38	0.39	0.41	0.24	0.31	1.00

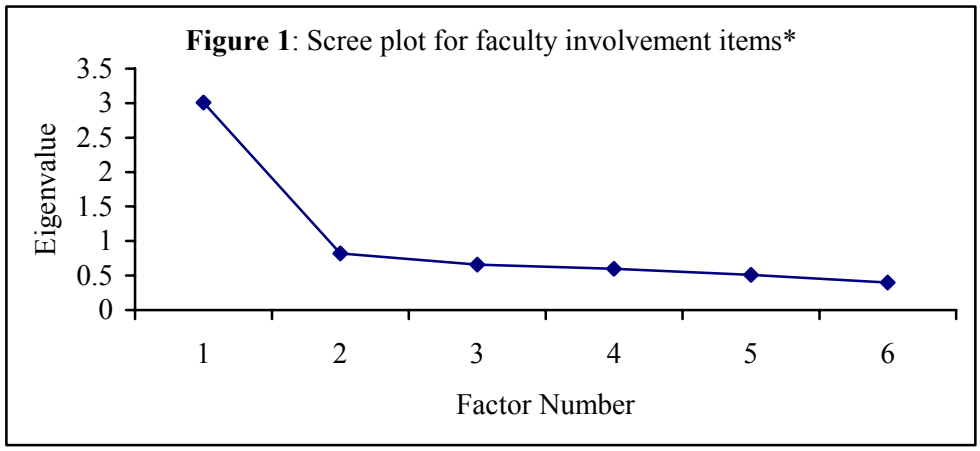
Assumption Checking. Table 3 shows the single factor solution for the faculty involvement item set. From this information we can begin to make a case that the faculty involvement items are unidimensional and do not violate the assumptions of IRT. Supporting a one factor solution as most appropriate, all factor loadings for the faculty involvement variables are quite high, ranging from .53 to .80. In addition, the ratio of the first to second eigenvalue is 3.69, a high ratio; this fact, combined with the fact that all but the first eigenvalues are quite small (and relatively similar in size), provide evidence that a one-factor solution is most appropriate (Hutten, 1980; Lord, 1980). Visually supporting the examination of eigenvalues, the

Scree plot in Figure 1 demonstrates an unmistakable "bend" or "elbow" with only one point above the elbow; this again suggests that a one-factor solution is most appropriate (Cattell, 1966).

Table 3
Factor Loadings and Eigenvalues For The Items Comprising Faculty Involvement Scales†

		Factor Loading	Eigen-values	Ratio of 1st to 2nd Eigenvalue
1	Freq: Interact with faculty outside class/office hours	0.58	1 3.01	3.69
2	Frequency: Asked a professor for advice after class	0.68	2 0.82	
3	Yes/No: Communicate regularly with your professors	0.80	3 0.66	
4	Satisfaction: Amount of contact with faculty	0.56	4 0.60	
5	Freq: Prof. provide advice about educational program	0.64	5 0.51	
6	Yes/No: Ever go to office hours	0.53	6 0.40	

†Extraction Method: Principal Axis Factoring, promax rotation; Polychoric correlation matrices used for analyses



*As computed from polychoric correlation matrix, see Table 2

Table 4 shows the residual correlation matrix created by subtracting the observed correlation matrix from the model-reproduced correlation matrix. Additional evidence supporting a one-factor (unidimensional) solution for the faculty involvement items is found in this table, as it demonstrates that a one-factor solution reproduced the observed correlations among the items well. After subtracting the reproduced from the observed correlations, the residuals among the faculty involvement items had a mean of .001 and a variance of .001. Further, most residual

correlations had absolute values less than .05, and none exceeded .07. These results not only argue for unidimensionality but also, as discussed above, provide evidence of "local independence," which is a critical assumption of IRT.

Table 4
Residual correlation matrix created (observed correlation matrix minus model-reproduced correlation matrix based on factor solution shown in Table 3)

		1	2	3	4	5	6
1	Freq: Interact with faculty outside class/office hours	0.00					
2	Frequency: Asked a professor for advice after class	0.00	0.00				
3	Yes/No: Communicate regularly with your professors	-0.01	-0.01	0.00			
4	Satisfaction: Amount of contact with faculty	-0.01	-0.05	0.06	0.00		
5	Freq: Prof. provide advice about educational program	-0.04	0.03	-0.01	0.05	0.00	
6	Yes/No: Ever go to office hours	0.07	0.03	-0.02	-0.05	-0.02	0.00

Step 2: Parameter Estimation

The parameters estimated by MULTILOG for the Faculty Involvement items are listed in Table 5.

Table 5
IRT parameters for faculty involvement items

	A	B1	B2	B3	B4	B5
Freq: Interact with faculty outside class/office hours	1.18	-1.17	0.16	1.19	2.21	3.60
Frequency: Asked a professor for advice after class	1.74	-1.21	1.36			
Yes/No: Communicate regularly with your professors	2.71	-0.90	1.10			
Satisfaction: Amount of contact with faculty*	1.20	-4.34	-2.76	-0.76	1.59	
Freq: Prof. provide advice about educational program	1.69	-0.87	1.48			
Yes/No: Ever go to office hours**	1.29	-2.24				

* "Can't rate" option coded as missing; ** Recoded from frequency of going to office hours

Step 3: Scoring

Original Score and Rescaled Score. Using the parameters in Table 5 and MULTILOG's scoring algorithm, each student in the 2008 YFCY dataset who answered at least one of the questions in the item pool was given a construct score for faculty interaction. The scores as obtained from MULTILOG ranged from -2.01 to 2.30, with a mean of 0.07 and a standard

deviation of 0.82. We rescaled the scores by multiplying each by 10 and adding 50, resulting in final score estimates that had a mean of 50.7 and a standard deviation of 8.2. See table 6 for a comparison of original and rescaled scores.

Table 6
Faculty Interaction Construct Scores

		Original Scale	Rescaled*
N	Valid	41047	41047
	Missing	70	70
Mean		0.07	50.68
Median		0.12	51.21
Std. Deviation		0.82	8.21
Minimum		-2.01	29.88
Maximum		2.30	72.96

*Rescaled score = Original Score*10 + 50

References

Allen, M. & Yen, W. (1979/2002). *Introduction to measurement theory*. Waveland Press: Long Grove, IL.

Astin, A. (1999). Student involvement: a developmental theory for higher education. *Journal of College Student Development, 40(5)*, 518-529. (Reprinted from Astin, A. (1984). Student involvement: a developmental theory for higher education. *Journal of College Student Personnel, 25,* 297-308).

Cattell, R. (1966). The scree test for the number of factors. *Multivariate Behavioral Research, 1(2)*, 245-276.

Clark, L.A. & Watson, D. (1995). Constructing validity: basic issues in objective scale development. *Psychological Assessment, 7(3)*, 309-319.

Cortina, J. (1993). What is coefficient alpha? An examination of theory and application. *Journal of Applied Psychology, 78(1)*, 98-104.

Dolan, C. (1994). Factor analysis of variables with 2, 3, 5 and 7 response categories: a comparison of categorical variable estimators using simulated data. *British Journal of Mathematical and Statistical Psychology,* 47, 309–326.

Embretson, S. & Reise, S. (2000). *Item response theory for psychologist*. Mahwah, NJ: Lawrence Earlbaum Associates.

Fox, J. (2009). Polycor: polychoric and polyserial correlations. R package version 0.7-7. http://cran.r-project.org/web/packages/polycor/

Gardner, P. (1995). Measuring attitudes to science: unidimensionality and internal consistency revisited. *Research in Science Education, 25(3)*, 283-289.

Hutten, L. (1980, April). *Some empirical evidence for latent trait model selection*. Paper presented at the annual meeting of the American Educational Research Association, Boston, MA.

Jöreskog, K. & Sorbom, D. (1989). *LISREL 7: A guide to the program and applications*. Chicago: SPSS, Inc.

Lord, F. & Novack, M. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Lord, F. (1980). *Applications of item response theory to practical testing problems*. New York: Earlbaum Associates.

McDonald, R. (1982). Linear versus nonlinear models in item response theory. *Applied Psychological Measurement,* 6, 379-396.

Olsson, U. (1979). Maximum likelihood estimation of the polychoric correlation coefficient. *Psychometrika,* 44 , 443-460.

Ostini, R. & Nering, M. (2006). *Polytomous item response theory models* . Quantitative Applications in the Social Sciences, Vol. 144. Thousand Oaks, CA: Sage Publications

Pascarella, E., Cruce, T., Umbach, P., Wolniak, G., Kuh, G., Carini, R., Hayek, J., Gonyea, R., Zhao, CM. (2006). Institutional selectivity and good practices in undergraduate education: how strong is the link? *The Journal of Higher Education, 77(* 2) , 251-285.

R Development Core Team (2009). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org.

Reise, S., Waller, N. & Comrey, A. (2000). Factor analysis and scale revision. *Psychological Assessment,* 12*(3)* , 287-297.

Revelle, W. (2009). psych: procedures for psychological, psychometric, and personality research. R package version 1.0-67. http://CRAN.R-project.org/package=psych

Russell, D. (2002). In search of underlying dimensions: the use (and abuse) of factor analysis in Personality and Social Psychology Bulletin. *Personality and Social Psychology Bulletin, 28(12)* , 1629-1646.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement, No.* 17 .

Tabachnick, B. & Fidell, L. (2007). *Using multivariate statistics* (5th ed.). Boston, MA: Pearson.

Thissen, D. Chen W.H., & Bock, D. (2002). MULTILOG, 7. Chicago: Scientific Software Incoroporated.