

## CIRP Freshman Survey: Reliability and Validity

HERI often receives questions from outside researchers asking about the ‘reliability’ and ‘validity’ of the CIRP surveys. The text below clarifies those two concepts and provides information about how they relate to the surveys conducted by the Cooperative Institutional Research Program.

**Reliability** refers to the *accuracy* of a given measurement. A reliable survey question, then, is one that yields similar results when administered repeatedly to similar samples or populations. Since the CIRP Freshman Survey has been administered for nearly forty years, it is possible to observe the stability of survey questions administered to different cohorts year after year. The vast majority of CIRP Freshman Survey questions exhibit a great deal of stability over time. In other words, in repeated trials the aggregate results remain very similar (allowing for modest fluctuations due to sampling error). Changes that are observed do not represent wild or random fluctuations, but can be linked to temporal trends or to real and meaningful exogenous shocks (the events of September 11th, for example). While nothing can be done to completely eliminate fluctuations caused by sampling error, several factors help ensure that such effects are minimized. Nearly 90 percent of the institutions in the CIRP Freshman Survey sample, for example, are repeat participants. This helps to ensure that the sample is highly consistent from year to year. Secondly, to correct for ‘response bias’ that occurs due to fluctuations in the annual sample, the data are stratified and weighted by institutional type and gender. For more information about calculating the margin-of-error for aggregate percentages from the CIRP Freshman Survey, please refer to the attached document, “The Precision of the Normative Data and Their Comparisons.”

**Validity**, by contrast, refers to whether a given survey question actually taps into the true underlying concept it attempts to measure. In other words, how well does the measure correlate with some unknown underlying ‘reality’? Factor analysis is one way for researchers to test the validity of certain constructs. By clustering related items together in scales or indices, for example, researchers can examine how well those related items ‘hold together’ in a statistical sense (this ‘scale reliability’ is most commonly measured using Chronbach’s Alpha or measured using factor loadings derived from factor analyses). While HERI has not performed in-depth factor analysis for every item on the CIRP Freshman Survey, several published works have investigated the matter. In What Matters in College, (Astin, 1992), for example, ‘student types’ were constructed based on

select items from the CIRP Freshman Survey (the leader, the scholar, the hedonist, etc). Astin found that the ‘student types’ held together quite well with most of the scale reliability coefficients in excess of .70. For more information, see also Astin’s Assessment for Excellence (Astin, 1991). Another article, ‘Linking Student Precollege Characteristics to College Development Outcomes: The Search for a Meaningful Way to Inform Institutional Practice and Policy’, by Jiali Luo and David Jamieson-Drake also contains several factors constructed using CIRP Freshman Survey items. The article can be found at: <http://airweb.org/page.asp?page=266>. HERI intends to perform a more in-depth factor analytical exploration of the CIRP Freshman Survey when time permits. However, at this time, we have not completed such analyses.

The information above refers mostly to the CIRP Freshman Survey which has the longest history and the highest participation rate and largest sample size of all of HERI's surveys. Our other surveys (the College Senior Survey and the Your First College Year survey) are administered on a much smaller scale and are not usually representative at the national-level. However, despite not being nationally representative, a large number of the questions on these two surveys are the same or very similar to the CIRP Freshman Survey. Therefore, it is safe to assume that most of the information provided above applies to these other surveys as well -- the primary difference being that the margins of error (as found in the attached document) are likely to be slightly larger due to the considerably smaller sample sizes.

## Appendix E

### The Precision of the Normative Data and Their Comparisons

A common question asked about sample surveys relates to the precision of the data, which is typically reported as the accuracy of a percentage “plus or minus x percentage points.” This figure, which is known as a confidence interval, can be estimated for items of interest if one knows the response percentage and its standard error.

Given the CIRP’s large normative sample, the calculated standard error associated with any particular response percentage will be small (as will its confidence interval). It is important to note, however, that traditional methods of calculating standard error assume conditions which, (as is the case with most real sample survey data), do not apply here. Moreover, there are other possible sources of error which should be considered in comparing data across normative groups, across related item categories, and over time. In reference to the precision of the CIRP data, these concerns include:

- 1) Traditional methods of calculating standard error assume that the individuals were selected through simple random sampling. Given the complex stratified design of the CIRP, where whole institutions participate, it is likely that the actual standard errors will be somewhat larger than the standard error estimates produced through traditional computational methods. In addition, while every effort has been made to maximize the comparability of the institutional sample from year to year (repeat participation runs about 90 percent), comparability is reduced by non-repeat participation and year-to-year variation in the quality of data collected by continuing institutional participants. While the CIRP stratification and weighting procedures are designed to minimize this institutional form of “response bias,” an unknown amount of non-random variation is introduced into the results.
- 2) The wording of some questions in the survey instrument, the text and number of response options, and their order of presentation have changed over the years. We have found that even small changes can produce large order and context effects. Given this, the *exact* wording and order of items on the survey instrument (which is produced as Appendix B) should be examined carefully prior to making comparisons across survey years.
- 3) Substantial changes in the institutional stratification scheme were made in 1968, 1971, 1975 and 2000. These changes resulted in a revision of the weights applied to individual institutions between 1966 and 2005. Stratification cell assignments of a few institutions may also change from time to time, but the scale of these changes and their effect on the national normative results are likely to be small in comparison to other sources of bias.

Since it is impractical to report statistical indicators for every percentage in every CIRP norms group, it is important for those who are interested to be able to estimate the precision of the data. Toward this end, Table E1 provides estimates of standard errors for norms groups of various sizes and for different percentages<sup>1</sup> which can be used to derive confidence interval estimates.

For example, suppose the item we are interested in has a response percentage of 18.7 percent among students at all nonsectarian four-year colleges (a normative group that is about 49,000 in size). First, we choose the column that is closest to the observed percentage 18.7 – in this case “20%”.<sup>2</sup> Next, we select the row closest to the unweighted sample size of 49,000 – in this case “50,000”. With a sample size of about 50,000 and a percentage that is close to 20, the estimated standard error would be .179.

To calculate the confidence interval at the 95% probability level, we multiply the estimated standard error by the critical value of *t* for the unweighted sample size (which, for all CIRP norms groups, will be equal to 1.96 at the .05 level of probability).<sup>3</sup> In this example, we would multiply the estimated standard error of .179 by 1.96, which yields .350. If we round this figure to a single decimal point we would then estimate our confidence interval to be 18.7 ± .4. In practical terms, this confidence interval means that if we were to replicate this survey using the same size sample, we would expect that the resulting percentage would fall between 18.3 percent and 19.1 percent 95 times out of 100.

Table E1  
*Estimated Standard Errors of Percentages for Norms Groups of Various Sizes*

Unweighted size of norms groups	Percentage										
	1%	5%	10%	15%	20%	25%	30%	35%	40%	45%	50%
500	.445	.975	1.342	1.597	1.789	1.936	2.049	2.133	2.191	2.225	2.236
1,000	.315	.689	.949	1.129	1.265	1.369	1.449	1.508	1.549	1.573	1.581
2,500	.199	.436	.600	.714	.800	.866	.917	.954	.980	.995	1.000
5,000	.141	.308	.424	.505	.566	.612	.648	.675	.693	.704	.707
7,500	.115	.252	.346	.412	.462	.500	.529	.551	.566	.574	.577
10,000	.099	.218	.300	.357	.400	.433	.458	.477	.490	.497	.500
15,000	.081	.178	.245	.292	.327	.354	.374	.389	.400	.406	.408
25,000	.063	.138	.190	.226	.253	.274	.290	.302	.310	.315	.316
50,000	.044	.097	.134	.160	.179	.194	.205	.213	.219	.222	.224
75,000	.036	.080	.110	.130	.146	.158	.167	.174	.179	.182	.183
100,000	.031	.069	.095	.113	.126	.137	.145	.151	.155	.157	.158
150,000	.026	.056	.077	.092	.103	.112	.118	.123	.126	.128	.129
275,000	.019	.042	.057	.068	.076	.083	.087	.091	.093	.095	.095

NOTE: Assumes simple random sampling.

<sup>1</sup>Calculated by  $\sqrt{\frac{x\%(100-x\%)}{N}}$ , where x is the percentage of interest and N is the population count from Table A3, column 2.

<sup>2</sup>Since the distribution of the standard errors is symmetrical around the 50 percent mid-point, for percentages over 50 simply subtract the percentage from 100 and use the result to select the appropriate column. For example, if the percentage we were interested in was 59, 100 – 59 percent yields 41, so we would use the column labeled ‘40%’.

<sup>3</sup>To calculate the confidence interval at the 99% probability level the critical *t* value is 2.56.