**Data Standardization Principles, Michael Hovland, University of Iowa**

First I want to thank Holly and Susan for their leadership of the Attributes group. I think we’ve made some good progress. Because I won’t be able to join the next attributes subcommittee meeting due to an earlier commitment, I wanted to take an opportunity to make a few comments that might be useful as we continue our discussions.

I have probably lived with this topic as long as anybody. My comments below reflect ways in which I have learned to view this process and ways that help me make decisions that help meet our objectives.

Scope of Our Work

It is important to remember that we are talking about standardizing data that only has low-stakes uses. In all cases we are talking about data that is gathered from vendors through search purchases and “hand-raiser” purchases. In other words, it is data that institutions receive at the pre-applicant end of the enrollment funnel. None of the data is used at the applicant stage. So, intended uses are for marketing, communications, and recruitment purposes. Thus, when we make decisions about fields like class rank and high school GPA, we are not looking at how these fields would be used for admissions applications or admissions decisions.

Simple is Better

Our decisions should reflect how the data is used in actual recruitment situations and shouldn’t be more complicated than what the uses dictate. A good example is Level of Parent Education. End-users are probably only interested in identifying for predictive purposes or communication purposes whether a student is first generation. So, rather than a list of ten or more levels of education, 4 or 5 might be sufficient, for example, less than high school, high school graduate, some college, 2-year degree, 4-year degree or higher.

Simple is always better for the data collectors as well. We are not making data standardization decisions in a vacuum. The decisions we make have to make sense for how vendors collect the data and for how institutions consume it.

How Data Is Collected Should Not Always be the Same as How Data Is Reported

Vendors are going to want to collect data in ways that students find intuitive. High School GPA is a good example. Some vendors like ACT collect and report GPA ranges in the form 3.75-4.0. Other vendors like College Board collect and report letter grades like B+. Still other vendors collect and report hybrid information like B (83-86). All of these approaches are problematic either at the collection stage or at the reporting stage. The ACT ranges don’t reflect the fact that many high schools have scales that go up to the 5s and 6s. The letter grade and hybrid letter grade/numerical range probably are intuitive for students in the data collection stage but don’t make much sense at all at the reporting stage. Colleges always store cumulative GPAs as decimal numbers and never as letter grades. Ranges are complicated but not impossible to store in databases. When we order search names, for example ACT 24-27, we have to store two separate fields (high and low) but in our more operational tables we store the mean score between 24 and 27. So, one possible recommendation would be to standardize on separate output fields depending on the vendor’s approach: Always report a number instead of a letter, and if you want to report a range, also report the mean in a separate field.

A related concern is that data must be consumable in standard ways. If I’m extracting a group of students for a mailing that have SAT scores > 1150 and HS GPAs >=3.0, I can do that easily when I’m dealing with numbers. If I have a field with letter grades representing GPAs, I can’t query the records where students have GPAs > B.

Reporting Decisions Should Reflect Campus Uses of the Data

Class rank is a good example here. An actual class rank requires a numerator and denominator; while this may be useful for admissions decisions, no search vendor is ever going to collect exact numerical ranks. So, ranges make sense. But what ranges? Again, these decision should be based on how campuses are likely to use the data. ACT uses four quartile ranges, but that doesn’t accommodate the institutions in states like Texas for example, that want to know if a student is top 10%. Most institutions are going to want to do everything they can to save money on search orders by ordering the names of students who are likely to be admissible. Our choice of class rank ranges should reflect practical differences related to likely admissibility.

Standardization Decisions Should Fit Database Requirements and Be Modifiable

Phone numbers are a good example. If we recommend reporting a single number in one field including country code, area code, prefix, and line number, for example, the number should be reported in such a way that an institution can reliably separate each of the component parts into separate fields of the data parts are stored separately in a system.

In short, most of the decisions we make will rely on logic and common sense and balance the separate needs for data collection and data reporting. And our specific focus should be primarily on data reporting. If College Board thinks it makes more sense to collect GPAs using letter grades, that’s fine. Just be cognizant of the fact that letter grades are not what colleges want to receive in output files.