University of Washington

IASystem™ Interpreting Reports

May 8, 2015

TABLE OF CONTENTS

INDIVIDUAL COURSE REPORTS

IASystem™ Individual Course Reports summarize student ratings of a particular class. They display a summary of numeric responses to evaluations conducted either online or on paper, and, for online evaluations, verbatim student comments. The following is a description of report content.

Header information

The name of the institution, college or school, and department of the course are printed at the top of each Individual Course Report, along with the academic term in which the class was taught.

Reports are also labeled with the course number and name, instructor name(s), whether the evaluation was conducted on paper or online, and the particular evaluation form used. The number of students who completed at least a portion of the evaluation form, class enrollment, and overall response rate are also displayed. Response rate is an important indicator of the reliability of the class ratings, and should be kept in mind when interpreting evaluation results.

General indices

Individual Course Reports present two general indices summarizing student ratings of the class.

OVERALL SUMMATIVE RATING

Four general items (described below) are included on most *IASystem*™ evaluation forms to provide a global rating of the class and instructor. They are rated from *Very Poor* to *Excellent* (0-5) and are summarized as a Combined Median. The items are:

The course as a whole was:
The course content was:
The instructor's contribution to the course was:
The instructor's effectiveness in teaching the subject matter was:

The Combined Median of the summative items is computed by first summing the numerical weights of all of the responses within each response category (e.g., all of the responses to *Excellent*, all of the responses to *Very Good*, etc.) across all four items. This provides a response array from which a median (ranging from 0-5) is calculated using the procedure described under *Computing Item Medians*, below. The resulting index is intended to be used in making high stakes summative decisions such as those relating to curricular development or faculty merit, promotion, and tenure. See *Using IASystem to make decisions*, below, for more information.

CHALLENGE AND ENGAGEMENT INDEX (CEI)

The Challenge and Engagement Index (CEI) provides an estimate of how challenging students found the class and how engaged they were in it. It is based on the combined response to four items included on most *IASystem*™ evaluation forms. The items are:

> *Relative to other college courses you have taken,*
> *The intellectual challenge presented was:*
> *The amount of effort you put into this course was:*
> *The amount of effort to succeed in this course was:*
> *From the total average hours [per week spent on the course], how many do you consider were valuable in advancing your education?*

IASystem™ transforms responses to each of these items to standard scores and calculates their average as described under *Computing the Challenge and Engagement Index*, below. The CEI correlates only modestly (~.25) with the Combined Median.

Item ratings

Individual Course Reports provide a rich perspective on student views by reporting responses to three categories of items.

- *Summative Items* are the first four items on most *IASystem*™ evaluation forms. These items are used to compute the global rating of the course and instructor, described above.
- *Student Involvement Items* are a set of items included on most *IASystem*™ evaluation forms to support computation of Adjusted Medians and the Challenge and Engagement Index.
- *Formative Items* relate to specific aspects of the course that instructors may want to change prior to the next iteration of the course. Responses to *Standard* and *Instructor-Added Formative Items* are reported separately.

Responses to individual items are reported in several ways: as frequency distributions, average (median) ratings, and either a) deciles or b) adjusted medians and relative ranks. Note that item text is not provided for *Instructor-Added Formative Items* added to paper evaluation forms; instructors should retain a copy of these items to assist in interpreting results.

FREQUENCY DISTRIBUTIONS

The total number of students who responded and the percentage of those students who selected each response choice are displayed for each item. Frequency distributions allow faculty to identify

unusual patterns of response. Instructors sometimes express the concern that evaluations may be completed primarily by students who feel strongly positive or strongly negative toward a course. When this is the case, the frequency distribution will be bi-modal.

ITEM MEDIANS

Individual Course Reports display average ratings in the form of item medians. Although means are a more familiar type of average than medians, they are less accurate in summarizing student ratings. Distributions of course evaluation item ratings tend to be strongly skewed. That is, most of the ratings are at the high end of the scale and trail off at the low end. The median indicates the point on the rating scale at which half of the students selected higher ratings, and half selected lower.

To interpret median ratings, compare the value of each median to the respective response scale. For example, a median of 4.5 on Items 1-4 means that the average rating is half-way between *Very Good* and *Excellent*. *IASystem*™ utilizes several different rating scales:

Excellent	Very Good	Good	Fair	Poor	Very Poor
5	4	3	2	1	0

Strongly Agree		Somewhat Agree	Somewhat Disagree		Strongly Disagree
6	5	4	3	2	1

Always, Much Higher, Very Much, Great		About Half, Half of the Time, Average, Moderate, Average			Never, Much Lower, Not at All, None
7	6	5	4	3	2	1

Note that for items relating to course workload, the median has been divided by credit hours to allow comparisons across classes.

STATISTICAL ADJUSTMENT OPTIONS

Deciles. Institutions may choose to display either deciles or a combination of adjusted medians and relative ranks to assist in interpreting course evaluation results. *Decile ranks* provide a means to compare the median rating of a particular item to ratings of the same item in all other classes within the college/school and across the institution. Values range from 0 (lowest) to 9 (highest); the lowest 10% of item medians are assigned a decile rank of 0, item medians above the bottom 10% and below the top 80% are assigned a decile rank of 1, etc. Note that because average ratings

tend to be high, a *Good* rating may have a low decile rank. *IASystem*™ restandardizes decile ranks annually for each institution based on ratings from the previous two academic years. Decile ranks are shown only for items for which there are sufficient data.

Adjusted medians. Institutions may choose to have *Adjusted Medians* displayed on Individual Course Reports in lieu of *Deciles*. Research has shown that student ratings may be somewhat influenced by factors such as class size, expected grade, and reason for enrollment. In particular, ratings may be lower for a) large classes, b) classes in which a high proportion of students expect low grades, and c) courses taken as a requirement rather than an elective. To control for these effects, *IASystem*™ analyzes institutional data to determine the strength of the observed relationships and applies a corrective formula to compute *Adjusted Medians* for the four *Summative Items* and the combined global rating. The formula is recalculated annually for each institution based on ratings from the previous two academic years as described under *Computing Adjusted Medians*, below.

Relative rank. Individual Course Reports that display adjusted medians for *Summative Items* display *Relative Rank* for *Formative Items*. These rankings are specific to the particular class evaluated. Relative ranks are computed by standardizing and rank ordering the median ratings of the items. Scores are standardized by subtracting the item median from the overall institutional item median and dividing by the standard deviation across all courses. The standardized scores are then rank ordered, with 1 being the highest ranked item with respect to that particular course. These ratings are intended to serve as a guide to direct instructional improvement efforts, with the top ranked items (1, 2, 3, etc.) representing the strongest areas and the lowest ranked items perhaps in need of additional focus.

Student comments

Responses to open-ended questions are provided as a separate report for evaluations conducted online. Colleges and schools within the institution determine whether these comments can be viewed by departmental coordinators and administrators. Comments are not available online for evaluations conducted on paper. Paper comment sheets should be collected by coordinators and sent to faculty after grades have been posted.

ADMINISTRATIVE REPORTS

Administrative reports provide information to department chairs and college deans to assist in curricular planning, faculty development, and personnel decisions.

High/Low Report

The High and Low Rated Courses and Instructors report supports curricular development by alerting administrators to courses that regularly receive especially high or low ratings. It also assists administrators in identifying faculty whose teaching is particularly strong, as well as instructors who may need additional support in their teaching. This report is especially useful when generated at the end of each academic term.

Highly rated faculty may be eligible for teaching awards or particularly able to help build instructional quality within the department. Chairs may wish to ask these faculty to collaborate in developing shared curricula, create and deliver teaching workshops, provide peer review for instructors coming up for promotion or tenure, and/or serve as teaching mentors. The report also identifies faculty who could benefit from additional departmental support to assist them in improving their teaching.

HEADER INFORMATION

The name of the institution, college or school, and department of the course are printed at the top of each High and Low Rated Courses and Instructors report, along with the academic terms in which the classes were taught.

EVALUATION GROUPINGS

Evaluation results are grouped into four sections: Highest Rated Faculty, Lowest Rated Faculty, Highest Rated TAs (teaching assistants), and Lowest Rated TAs. Grouping is based on the Combined Adjusted Median of the four summative evaluation items. The “highest” evaluations are those with a value greater than or equal to 4.7 (close to *Excellent*). Evaluations classified as “lowest” have a value less than 3.0 (less than *Good*).

RESULTS DISPLAYED

For each course/instructor combination, the report displays the course name and number, and instructor name and rank. Additional information includes course enrollment, evaluation response rate, and whether the evaluation was conducted online or on paper. Four summaries of evaluation results are reported for each course. The Combined Median, Adjusted Combined Median, and CEI

have been described above. They are reported for all evaluations. Student response to a fourth item (scaled *Excellent* to *Very Poor*, 5-0) is also reported for evaluations using forms that include this item:

Amount you learned in the course was:

Ratings Summary

The Ratings Summary report provides an overall view of evaluation results within a particular academic unit (department, college/school, or institution). It has been created to support annual program review, but can be generated for any time period.

HEADER INFORMATION

The name of the institution, college or school, and department are printed at the top of each Ratings Summary report, along with the academic terms in which the classes were taught.

RESULTS DISPLAYED

The Ratings Summary report summarizes student response to a selected set of items found on all evaluation forms. The combination of the four summative items is reported, along with two of those items and six of the student engagement items. The individual items reported are:

The course as a whole was:
The instructor's effectiveness in teaching the subject matter was:
The Combined Median
Relative to other college courses you have taken,
    Do you expect your grade in this course to be:
    The amount of effort to succeed in this course was:
On average, how many hours per week have you spent on this course?
What grade do you expect in this course?

Item responses are reported by instructor rank and course level (lower level course, faculty; lower level course, TA; upper level course; graduate level course) and total. Specific statistics reported are the number of evaluations in each category, the mean and standard deviation of the Combined Medians, and the mean and standard deviation of the Combined Adjusted Medians.

Evaluation List

The Evaluation List report displays all evaluations conducted within a particular academic unit (department, college/school, or institution) during a particular time period. It was designed to accompany the Ratings Summary report, but can be used independently as well.

HEADER INFORMATION

The name of the institution, college or school, and department of the course are printed at the top of each Evaluation List report, along with the academic terms in which the classes were taught.

RESULTS DISPLAYED

The Evaluation List report details all evaluations conducted within the specified time period. Entries are listed alphabetically by instructor name and ordered, within instructor, by course name and number. The report shows the academic term of the class, the number of credits, the number of enrolled students, the number of students who responded to the evaluation and the response rate. The evaluation form used and whether the evaluation was conducted online or on paper are also shown. Evaluation results are reported in the form of the Combined Median.

COMPUTING MEDIANS, ADJUSTED MEDIANS, AND THE CEI

Medians

Medians are a measure of central tendency that indicate the point on the scale dividing a distribution of scores or ratings evenly in half; half of the scores fall above the median, and half fall below. *IASystem*™ reports item medians rather than means because they more accurately represent student ratings of each item. Medians are computed to one decimal place by interpolation, and, for most items, higher medians reflect more favorable ratings.

MEASURES OF CENTRAL TENDENCY

The three measures of central tendency used to describe distributions of scores are the mean, median, and mode. Each has its own particular advantages and disadvantages depending on the shape of the score distribution. The **mean** is the most familiar and is the arithmetic average, calculated by adding up all the scores and dividing by the total number of scores. The **median** is the point on the scale that divides the distribution of scores in half (half of the scores fall above the median and half fall below). The **mode** is simply the score that occurs most frequently. Note that both the mean and the median are points on a scale and are found by computation; they aren't necessarily whole numbers.

If the score distribution is symmetrical, the mean, median, and mode are identical and fall in the middle of the scale. When distributions are not symmetrical, these three measures take on different values. Because of the way they're computed, means are influenced by extreme scores whereas medians are not. If a distribution is skewed, the mean is pulled out toward the tail of the distribution, while the median remains in the middle. Course ratings tend to be left-skewed, and for this reason *IASystem*™ average ratings are reported in the form of item medians.

The computation of *IASystem*™ medians is based on the method described by Guildford (1965)[1] and illustrated below. You may recognize this method as that used most commonly for calculating the median of grouped data. This method represents the actual ratings more precisely than does the "ordinal" median computed using un-grouped data.

COMPUTATIONAL EXAMPLE

In our example, 160 students rated a single item. The scale is 0-5 (*Very Poor* to *Excellent*) and the mean is 3.76. The median is the point on the scale that divides the distribution into halves, with 80 scores above and 80 scores below. As shown in Table 1, the scores don't divide themselves evenly

[1] Guilford, J.P. (1965). *Fundamental Statistics in Psychology and Education*, New York: McGraw-Hill, pp. 49-55.

into two groups, and the median would fall somewhere in the interval 4. The lower and upper limits of this interval are 3.5 and 4.5, respectively, and the exact value of the median is determined by the process of interpolation. In this process, the 74 scores are 'spread evenly' along the interval, and the median is located proportionately above the lower limit of 3.5 or below the upper limit of 4.5.

Table 1. Sample distribution of course ratings

Rating	Frequency	Cumulative Frequency	
0	1	1	54 cases below the median
1	1	2	
2	12	14	
3	40	54	
4	74	138	74 cases within the interval containing the median
5	32	160	32 cases above the median

INTERPOLATING UP FROM THE LOWER LIMIT

The formula to compute the median by interpolating up from the lower limit is:

$$L_m + I_m ((N / 2 – cf) / f_m)$$

Where: L_m = lower limit of the interval containing the median
I_m = the width of the interval containing the median
N = total number of scores
cf = cumulative frequency
f_m = number of scores within the interval containing the median

This is illustrated in the following steps.

Step	Result
Identify the lower limit of the interval containing the median	3.5
Find the width of the interval	4.5 - 3.5 = 1.0
Determine the number of scores needed above the lower limit	80 – 54 = 26
Determine the proportion of the interval above the lower limit	26 / 74 = .35
Convert the proportion to scale units	35 * 1.0 = .35
Find the scale value of the median	3.5 + .35 = 3.85

INTERPOLATING DOWN FROM THE UPPER LIMIT

Interpolating down from the upper limit will give the same median value as interpolating up from the lower limit, as shown below.

Step	Result
Identify the upper limit of the interval containing the median	4.5
Find the width of the interval	4.5 - 3.5 = 1.0
Determine the number of scores needed below the upper limit	80 – 32 = 48
Determine the proportion of the interval below the upper limit	48 / 74 = .64
Convert the proportion to scale units	.64 * 1.0 = .64
Find the scale value of the median	4.5 - .64 = 3.85

Note: Although we have reported medians to two decimals in this example to illustrate the method of computation, they are reported to only a single decimal on summary reports.

Computing adjusted medians

The goal of adjusting student ratings is to remove the effect of known biasing factors. Early research at the University of Washington found a significant relationship between students' expected course grades and their ratings of the course.[2] Other factors identified at the UW and elsewhere are student reason for enrollment and class size. *IASystem*™ uses a multiple regression approach to identify the existence and strength of biases with respect to these factors at each individual institution. Analyses are carried out annually and are based on ratings from the previous two years. These procedures enable *IASystem*™ to create and regularly update institution-specific formulae to adjust the medians of the four summative items as well as the Combined Median. The computational definitions of the adjustment variables are as follows:

- RG (Relative Grade) = class mean of the following item: *Relative to other college courses you have taken, do you expect your grade in this class to be:* (Much Lower to Much Higher, 1-7) ;
- ER (Enrollment Reason) = percentage of students taking the course in their major, minor, as an elective, or other (as opposed to as a program or distribution requirement); and
- LS (Log of Class Size) = the natural log of class size

[2] Greenwald, A.G. and Gillmore, G.M. (1997). Grading leniency is a removable contaminant of student ratings. American Psychologist, 53, 1209-1217.
Greenwald, A. G. and Gillmore, G. M. (1997). No pain no gain? The importance of measuring course workload in student ratings of instruction. Journal of Educational Psychology, 89 (4), 743-751.

UNIVERSITY OF WASHINGTON EXAMPLE

The following example is based on analysis of ratings of 21,000 classes conducted over two years at the University of Washington. Simple correlations between the item, *The course as a whole was*, and three adjustment variables are shown below.

Table 2. Zero-order correlations between *The course as a whole was* and three adjustment variables

r	Adjustment variable
.34	Relative Grade (RG)
.21	Enrollment Reason (ER)
-.25	Log of Class Size (LS)

As the table shows, students tend to award higher ratings if they are expecting a high grade in the course; if they are taking a class in their major, minor, or as an elective rather than as a program or distribution requirement; and if the class is relatively small. When these three variables were used to predict the item median, the resulting regression equation was as follows:

$$\text{Predicted item median} = 3.148 + .247*RG + .00347*ER - .148*LS$$

This can be re-written in a form to compute the adjusted median:

$$\text{Item median} - [3.148 + .247*RG + .00347*ER - .148*LS - 4.134]^{3}$$

In this particular example, the regression equation explained 20% of the total item variance, and the correlation between the adjusted and unadjusted medians was .89.

Table 3 translates this formula into the specific adjustments made for several values of each adjustment variable. The adjustments are also represented visually in Figures 1-3.

Table 3.

Relative Grade (RG)		Enrollment Reason (ER)		Class Size (LS)	
Average	**Adjustment**	**Percentage**	**Adjustment**	**Size**	**Adjustment**
3.5	+.35	25%	+.17	5	-.24
4.0	+.22	40%	+.12	10	-.14
4.5	+.10	60%	+.05	25	-.00
5.0	-.02	80%	-.02	50	+.10

[3] In this equation, two constants have been entered separately representing the intercept (2.487) and the grand mean (3.8829).

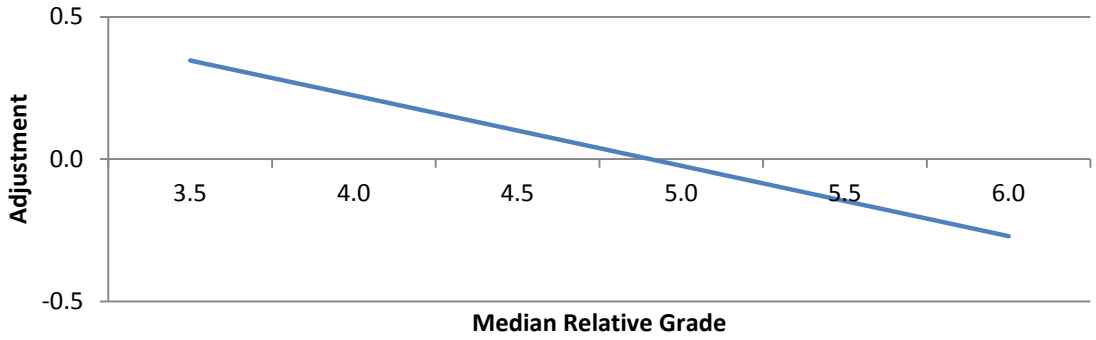

Figure 1. Adjustment to Item 1 Median based on Relative Grade

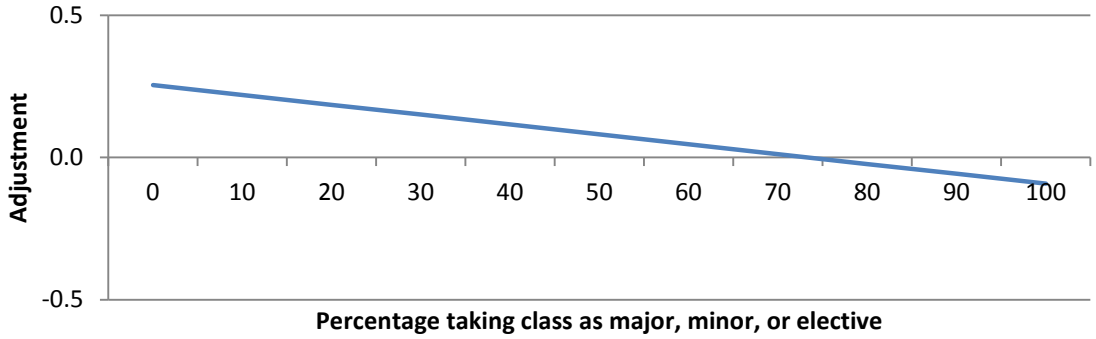

Figure 2. Adjustment to Item 1 Median based on Enrollment Reason

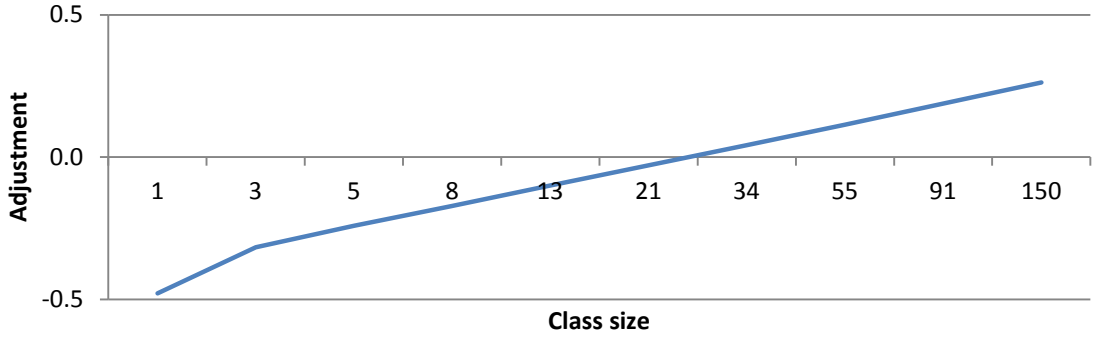

Figure 2. Adjustment to Item 1 Median based on Class Size

Computing the Challenge and Engagement Index (CEI)

The Challenge and Engagement Index (CEI) combines student responses to several items to provide an overall estimate of how academically challenging students found the class and how engaged they were in it. Development of the index was sparked by analyses[4] of *IASystem*™ data suggesting that:

- Students put more effort into classes that demand more effort for them to be successful.
- Students tend to prefer more challenging classes over less challenging classes.
- The widely held belief that assigning students more work will lead to lower student ratings is not true in and of itself.
- All faculty are not equally demanding; there are considerable differences across faculty in the amount of time students devote to their courses.

The four items comprising the CEI scale are:

Relative to other college courses you have taken, the
intellectual challenge presented was:
amount of effort you put into this course was:
amount of effort to succeed in this course was:
For the total average hours [per week spent on the course], how many do you consider were valuable in advancing your education?

Because these items utilize different rating scales, *IASystem*™ converts student responses to standard scores before creating a combined index. The first three items range from *Much Lower* to *Much Higher* (1-7) whereas twelve different categories are used for the last item. After converting the items to standard scales, *IASystem*™ computes a Combined Median using a scale of 1-7, following the procedure described under *Computing Item Medians*, above.

The CEI represents a somewhat different aspect of the course than does the Combined Median; it is only modestly correlated (.25) with the four summative items and their Combined Median.

[4] Gillmore, G.M. (2001). What student ratings results tell us about academic demands and expectations. OEA Report 01-02. uw.edu/oea/pdfs/reports/OEAReport0102.pdf

USING *IASYSTEM*™ TO MAKE DECISIONS

IASystem™ provides feedback to instructors for the purpose of course improvement (formative decision-making) and to faculty and administrators to inform curricular development and merit, promotion, and tenure decisions (summative decision-making). To support the quality of decision-making at the level of the classroom and department, *IASystem™* utilizes several different strategies, regularly assesses item quality as reflected in reliability indices, and recommends specific interpretive guidelines for making both *formative* and *summative* decisions.

Strategies to ensure data quality

IASystem™ employs several strategies to ensure high quality data and encourage appropriate use of information for both formative and summative decisions.

STRUCTURE OF RATINGS FORMS

Different types of items are included on *IASystem™* evaluation forms to support different evaluative functions.

- The four *summative* items are included on most *IASystem*™ evaluation forms provide a global rating of the class and instructor. These items are combined into a single index to be used in making high stakes decisions such as those relating to curricular development or faculty merit, promotion, and tenure. Combined Medians from multiple classes can be averaged to compare one course to another, ratings of an individual instructor to a set criterion, or to look for change in ratings over time.
- *Student involvement* items are also included on most *IASystem*™ evaluation forms. This item set includes items used to adjust bias in ratings of summative items and items used to create the Challenge and Engagement Index.
- *IASystem*™ evaluation forms differ from one another in items focused on specific aspects of various types of courses. Responses to these items enable instructors to make *formative* changes to their courses prior to the next time the course is offered.

COMPUTED INDICES

High stakes decisions relating to courses or instructors should not be made on the basis of ratings of individual items, but on combined data. *IASystem*™ reports a Combined Median for each evaluation, providing a single, overall summary of student assessment of the course. *IASystem*™ also reports a second index reflecting the level of student engagement in the course. The Combined Median and Challenge and Engagement Index represent different, but slightly related, aspects of the course.

BIAS CONTROL

Analysis of student ratings data reveals that student course ratings are influenced by several factors. The best known of these is expected grade in the course, student reason for taking the class, and class size. *IASystem™* corrects for observed bias by computing and reporting Adjusted Medians based on institution-specific algorithms. (See *Computing Adjusted Medians*, above.)

NORMATIVE COMPARISONS

Faculty and administrators may want to develop specific criteria to assist in decision-making, particularly when making high-stakes summative decisions. *IASystem™* provides both *relative* and *absolute* norms as standards for comparisons.

When using *relative* standards, decisions are made by comparing ratings of a single item, course, or combination of courses to ratings of other items or courses. *IASystem™ decile ranks* compare the median ratings of all summative and formative items, as well as the Combined Median, to the same ratings in other classes within the college/school and across the institution.[5] *Relative ranking* of formative items uses normative comparisons to rank order student ratings of specific areas of a course.

Absolute standards entail comparison of item or course ratings against a pre-set criterion, such as a median rating of a particular value. For these types of comparisons, item reliability is particularly important as described below.

Item reliability

A key measure of the quality of student ratings forms is the reliability of individual item ratings. Two types of reliability estimates are computed for *IASystem™* items: inter-rater and inter-class reliability.[6]

Formative decisions generally relate to modifications to a course or instruction and usually are made based on evaluation of an individual class. For this type of decision, inter-rater reliability coefficients are the appropriate index, representing the degree of confidence faculty can have in making such changes. Reliability indices increase with the number of cases, and adequate

[5] *IASystem™* does not support normative comparisons between colleges and universities due to the high variance in institutional cultures.

[6] Gillmore, G. M. (2002). Drawing Inferences about Instructors: The Inter-Class Reliability of Student Ratings of Instruction. *OEA Reports*, *00-02*. www.washington.edu/oea/pdfs/reports/OEAReport0002.pdf

reliability ($r > .70$)[7] for formative decision-making is obtained with class sizes of seven students or more for both summative and formative items as shown in Figure 4.

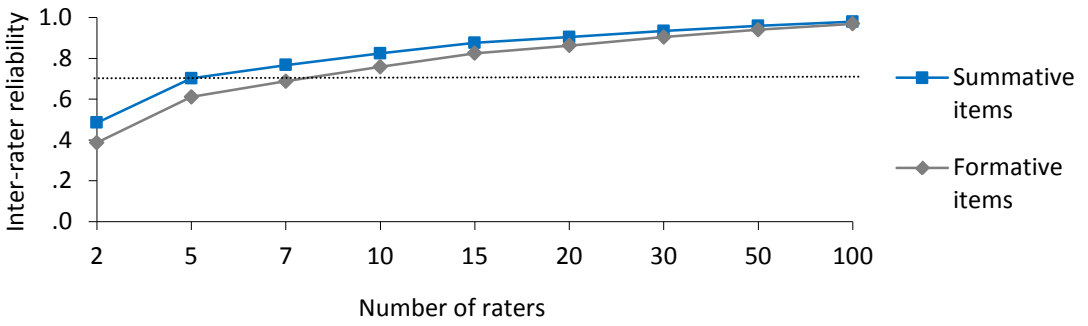

Figure 4. Inter-rater reliability of *IASystem™* items for formative (single-class) decision-making

Summative decisions, particularly those relating to instructor merit, promotion, and tenure, require a higher level of certainty than do formative decisions. Summative decision-making should be based on ratings combined over multiple classes and appropriate reliability estimates must reflect the added instructor effect. In this context, *IASystem™* utilizes the inter-class reliability coefficient, assessing the stability of ratings across classes, and adequate reliability for summative decision-making is obtained when the number of classes equals five or more as illustrated in Figure 5.

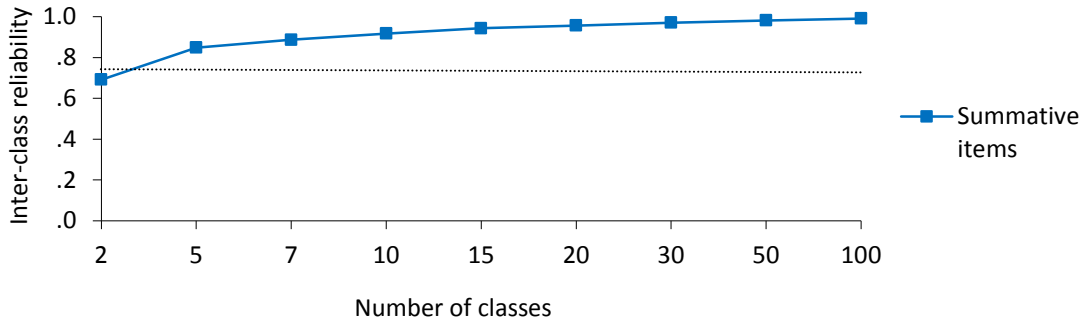

Figure 5. Inter-rater reliability of *IASystem™* items for summative (combined classes) decision-making

An important corollary of item reliability is the magnitude of the observed difference required for statistical significance when comparing average ratings against a criterion.[8] In making merit, promotion and tenure decisions, department chairs often compare average ratings for a particular instructor against a pre-set standard. For example, instructors may be considered eligible for merit

[7] Reliability coefficients range from 0.0 to 1.0, with higher numbers indicating more agreement among raters and lower coefficients indicating less agreement. As a general rule of thumb, low, medium and high reliability are referenced by coefficients of .00-.40, .40-.70, and .70-1.00, respectively.

[8] McGhee, D.E. (2002). Drawing inferences about instructors: Constructing confidence intervals for student ratings of instruction. *OEA Reports, 02-05*. www.washington.edu/oea/pdfs/reports/OEAReport0205.pdf

pay if their average combined summative rating for the previous three years is at least 3.0 ("Good"). Reliability analysis tell us that an observed difference of at least .3 is required for statistical significance, so in this example, instructors average ratings of at least 2.7 would be awarded merit pay.

Interpretive guidelines

Departments can maximize the validity of decision-making process by developing thoughtful, systematic, and well-articulated policies and practices, consistent with known strengths and limitations of ratings data.

DECISIONS RELATING TO COURSE IMPROVEMENT

With respect to course-improvement, instructors appropriately decide to modify aspects of their courses from one academic term to the next. These are not high stakes decisions, the items that are used are specific to the course format, changes are made based on the particular item content, and adjustments in the course are made continuously over time. As noted above, *formative* items show adequate inter-rater reliability for class sizes of seven or more students.

Interpretive guidelines for *formative* decisions:

- Judgements may appropriately be based on ratings of individual items.
- Judgements may appropriately be based on ratings of individual courses.
- Decisions should be made on ratings provided by at least 7 students.

DECISIONS RELATING TO MERIT, PROMOTION, AND TENURE

The decision-making process must be more rigorous for merit, promotion, and tenure decisions. Not only are these decisions high-stakes, but the issue of item reliability doesn't concern ratings of a single class, but the extent to which the ratings for an instructor are consistent across several classes. As noted above, *summative* items show adequate inter-class reliability when data are combined for five or more classes, and reliability is further increased by combining the four summative items into a single combined rating.

Interpretive guidelines for *summative* decisions:

- Judgements should be based on the Combined Median, rather than ratings of individual items.
- Judgements should be based on the combined rating of at least 5 courses.
- Decisions should require a minimum $\pm$ .3 difference when comparing average ratings for a particular instructor against a criterion.