Context Matters: Assessment in Large Research Institutions

IUPUI Assessment Institute

October 13, 2019

Timothy S. Brophy

Professor and Director, Institutional Assessment, Office of the Provost, University of Florida, Gainesville, Florida USA

Today's Goals

To introduce the elements of effective assessment systems in the context of sustained excellence

Examine Validity and Reliability Conceptually and in Practice

Review and Discuss Faculty Assessment Data Reports

Common Challenges for Sustaining Excellence in Assessment

Institutional size and scope

- Multiple colleges/departments
- Diverse programs - Certificate, Undergraduate, Graduate, and Professional
- Available personnel

Institutional consistency

- Outcomes
- Assessment reporting
- Cycles of planning and reporting

Common Challenges for Sustaining Excellence in Assessment

Management and Tools

- Faculty assessment resources
- Templates, guidelines
- Professional development for faculty

Honoring unit autonomy, disciplinary distinctions, and institutional requirements

Faculty comportment

Part 1: Developing an Assessment System in a Large University

Element 1: Define the System

What is an assessment system? How do you define the elements of the system?

What is an *Assessment System*?

The Assessment System is a coordinated and carefully designed set of processes and tools used by university accreditation and quality assurance staff, administrators, and faculty to submit, review, store, and access institutional effectiveness and academic program assessment plans and data reports

Definitions

- **Institutional effectiveness**
  - the systematic, explicit, and documented process of *measuring performance against mission in all aspects of an institution*
- **Academic Assessment**
  - the systematic, explicit, and documented process of *measuring achievement of student learning outcomes and goals for an academic program*

Element 2: Establish the Institutional Framework

What is your Purpose, Mission, and Vision for assessment at your institution?

Establish Purpose and Mission

Purpose – Why you exist

The *purpose* of institutional assessment is to support the university's mission by establishing, maintaining, and refining the university's institutional effectiveness and assessment processes.

Mission – What you do

The *mission* of institutional assessment is to lead the university's efforts in accreditation and institutional effectiveness, assessment support, and to maintain transparent communication with all relevant stakeholders.

Develop your Vision:
Your mission achieved with excellence

We envision our university as an institution where all units and academic programs contribute to the fulfillment of its mission by establishing goals and outcomes, regularly assessing these with valid, reliable measures, analyzing and interpreting the data collected, and using the results for continuous improvement.

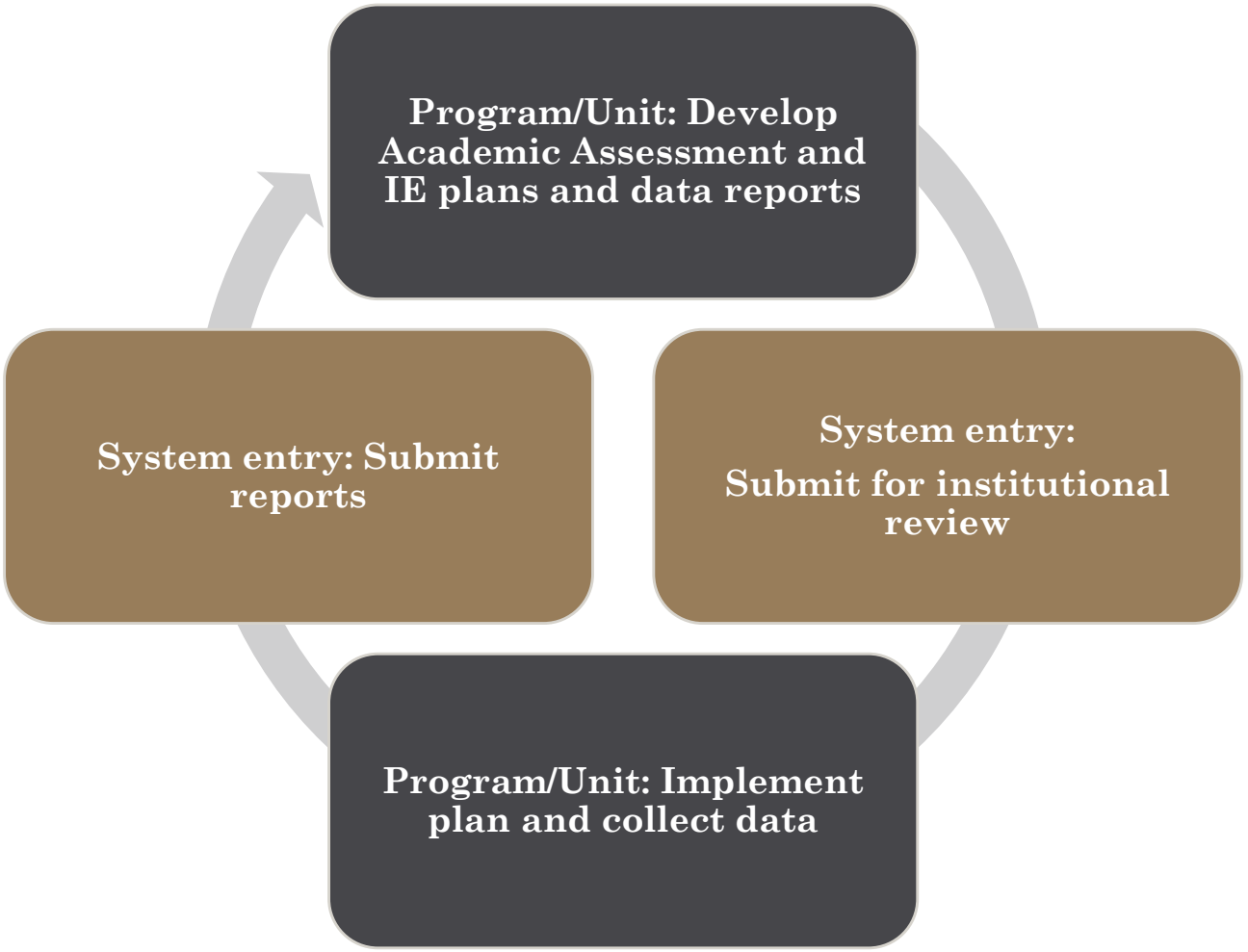
Element 3: Determine the System Inputs

What to enter, and when to do it

Data Sources

- **System Inputs** are those data, reports, and other information that are entered into the system
- Some examples:
  - Accreditation planning and reporting documents
  - New assessment plans
  - Modifications to existing Academic Assessment Plans
  - Assessment data reports
- There should be an **institutional cycle** for inputting this information

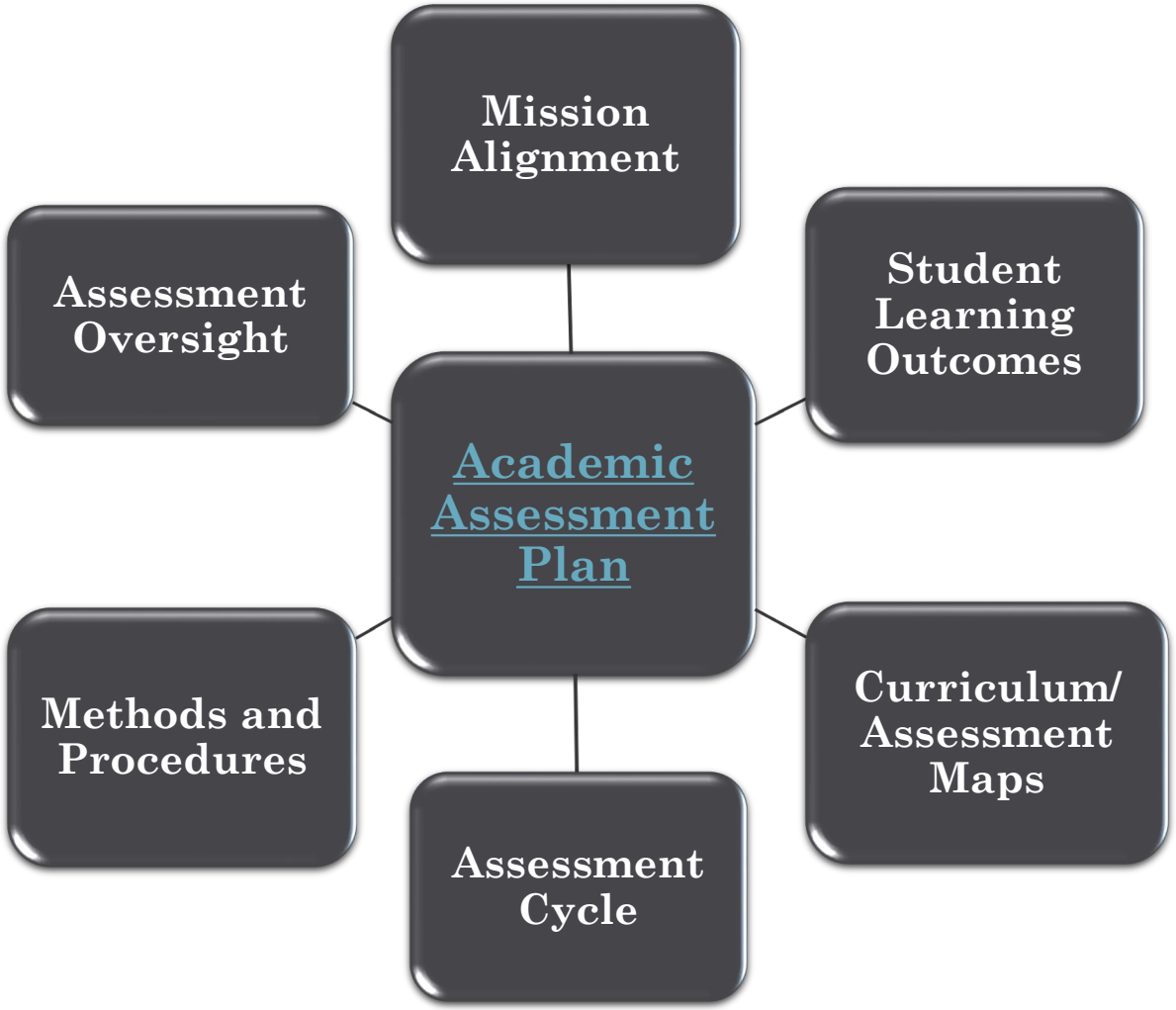
Planning Timeline/Cycle



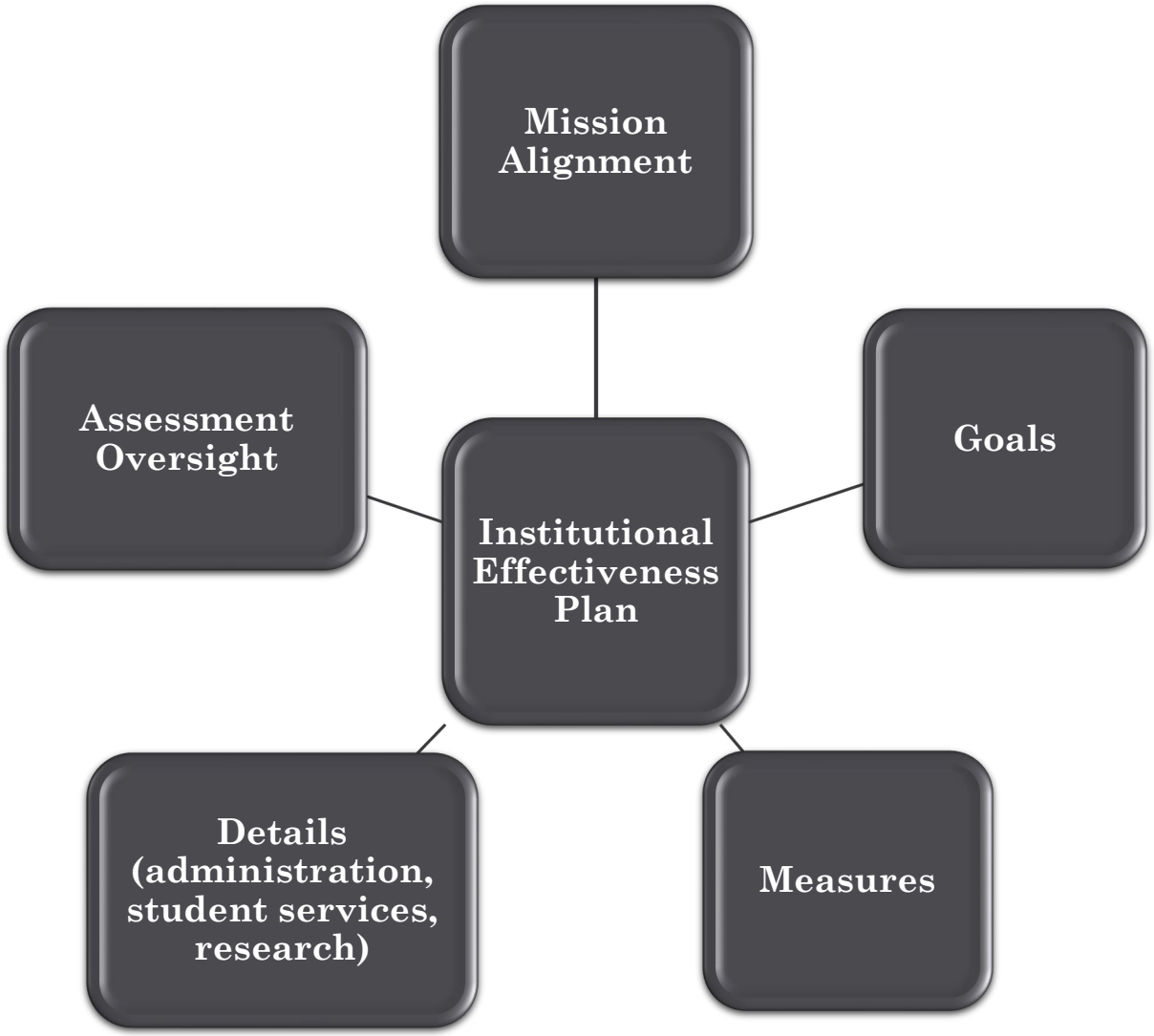
Element 4: Determine the Plan and Report Structure

How will your institution's assessment plans and reports be structured?

Assessment Plans for Academic Programs



Effectiveness Plans for Administrative Units



Data Report Components

Results for each of the Student Learning Outcomes and/or Goals

Use of Results for program/unit improvement is reported holistically

Element 5: Establish the System Processes

How an assessment system works

Faculty oversight

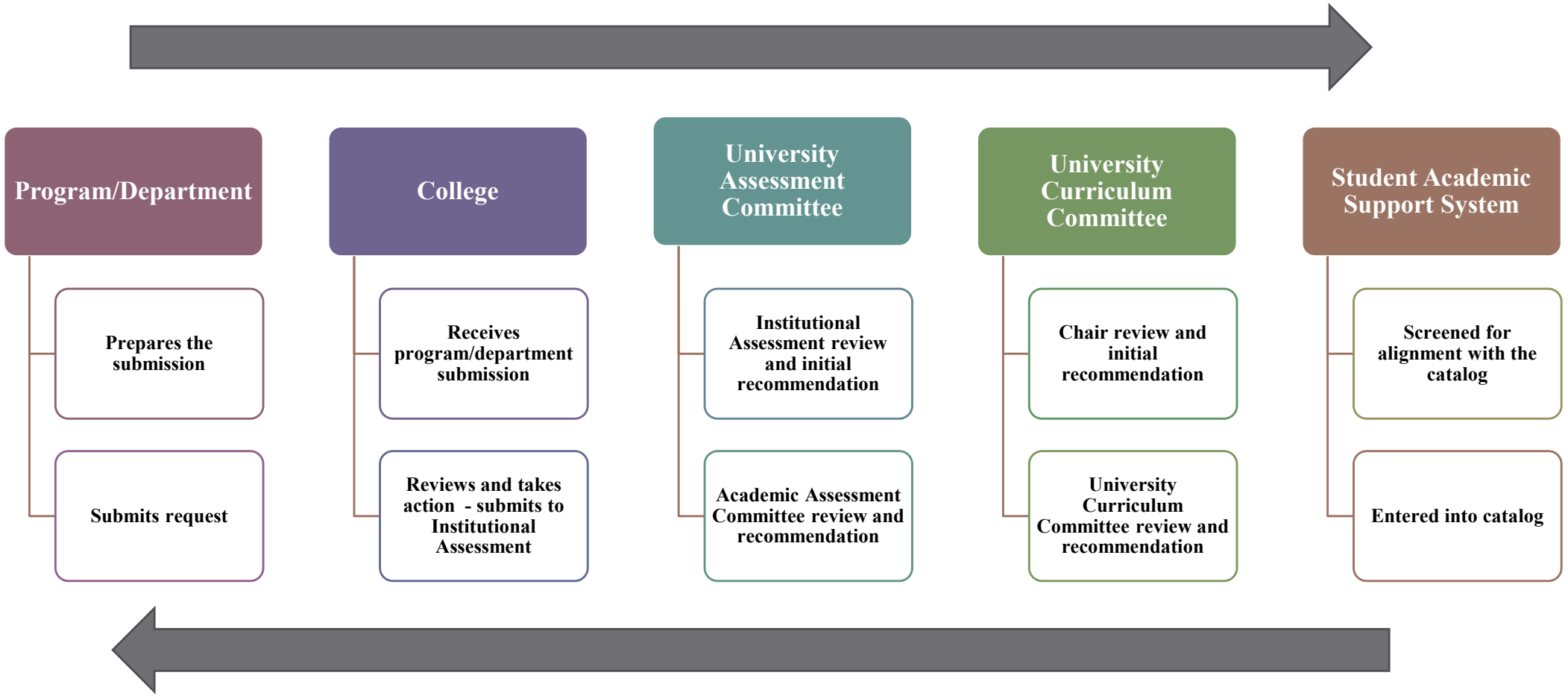
Institution-level committee of faculty, staff, liaisons

What they do:

Review and approve Academic Assessment Plans, including Student Learning Outcomes

Improve the efficiency of Institutional Assessment processes

Sample Multi-Stage Approval Process



Committee Review

Student learning outcomes – to ensure they meet your institution's expectations

- Student Learning Outcomes reflect the curriculum, the discipline, and faculty expectations; as these elements evolve, learning outcomes change.
  - *Recent* – the outcome reflects current knowledge and practice in the discipline.
  - *Relevant* – the outcome relates logically and significantly to the discipline and the degree.
  - *Rigorous* – the degree of academic precision and thoroughness that the outcome requires to be met successfully.
- Distinguish outcomes from outputs
- Distinguish outcomes from program goals
- Ensure that outcomes are measurable and valid for the SLO

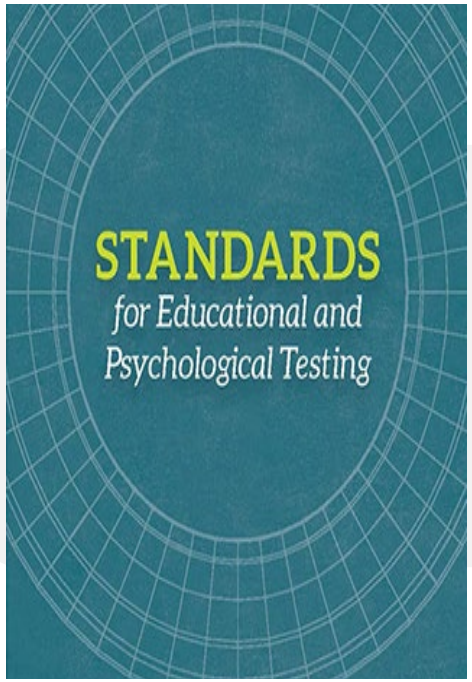
Communication

- Consider a *distributed leadership* model
- At a unit level appropriate for your institution, assign one person to be the contact for your office
- You communicate with them; they communicate to their faculty and administration
- These individuals can also meet as a group when needed

Element 6: Validity, Reliability, and Fairness

How do you address this at the institutional level?

Validity



Validity is **"a unitary concept – it is the degree to which all the accumulated evidence supports the intended interpretation of test scores for the proposed use."**

APA/AERA/NCME, Standards for Educational and Psychological Testing, 2014.

Checking for Validity at the Institutional level

Assessment staff and the relevant committee of faculty should review all plans and data reports

All measures of goals and SLOs are reviewed to ensure that they lead to data pertinent to the goal or outcome (validity)

If there are questions, the plan or report should be returned for modification or clarification

Reliability/ Precision and Fairness

In its broadest sense, "reliability refers to the consistency of scores across replications of a testing procedure…this is always important…and the need for precision increases as the consequences of decisions and interpretations grow in importance."

Fairness has to do with "equitable treatment of all test takers, lack of measurement bias, equitable access to the constructs as measured, and validity of individual test score interpretations for their intended uses."

- *APA/AERA/NCME, Standards for Educational and Psychological Testing, 2014.*

Checking for Reliability and Fairness at the Institutional Level

Reliability and Fairness of Student Learning Outcome assessments is the responsibility of the academic program faculty – they develop and administer the assessments

Faculty have access to the built-in reliability functions of our Learning Management System (Canvas) – they can program the Learning Management System to collect data on their program SLOs

Element 7: System Outputs

What does the assessment system produce?

Student Learning Outcomes Approvals

Program leaders are informed via email on any actions taken by the assessment committee

Options are:

- Approve
- Comment
- Conditionally Approve
- Table (rarely used)
- Recycle
- Denied

Constructive feedback

Provide feedback on all data reports and request modifications if needed

Allow a reasonable length of time for the modifications to be completed

At the University of Florida, our most common requests are to:

- report improvement actions as a decision made based on the review of results, in the past tense
- Remove any future tense phrases in the improvement actions

Examples of Feedback

- Art History (PhD) (program goal and SLO report)
  - *Excellent data summaries and documentation. PG1, PG 2, PG3, PG4, PG5, PG6, SLO1, SLO2, SLO3, SLO4 - slightly revise Use of Results to read as a decision made based on the use of results (past tense).*

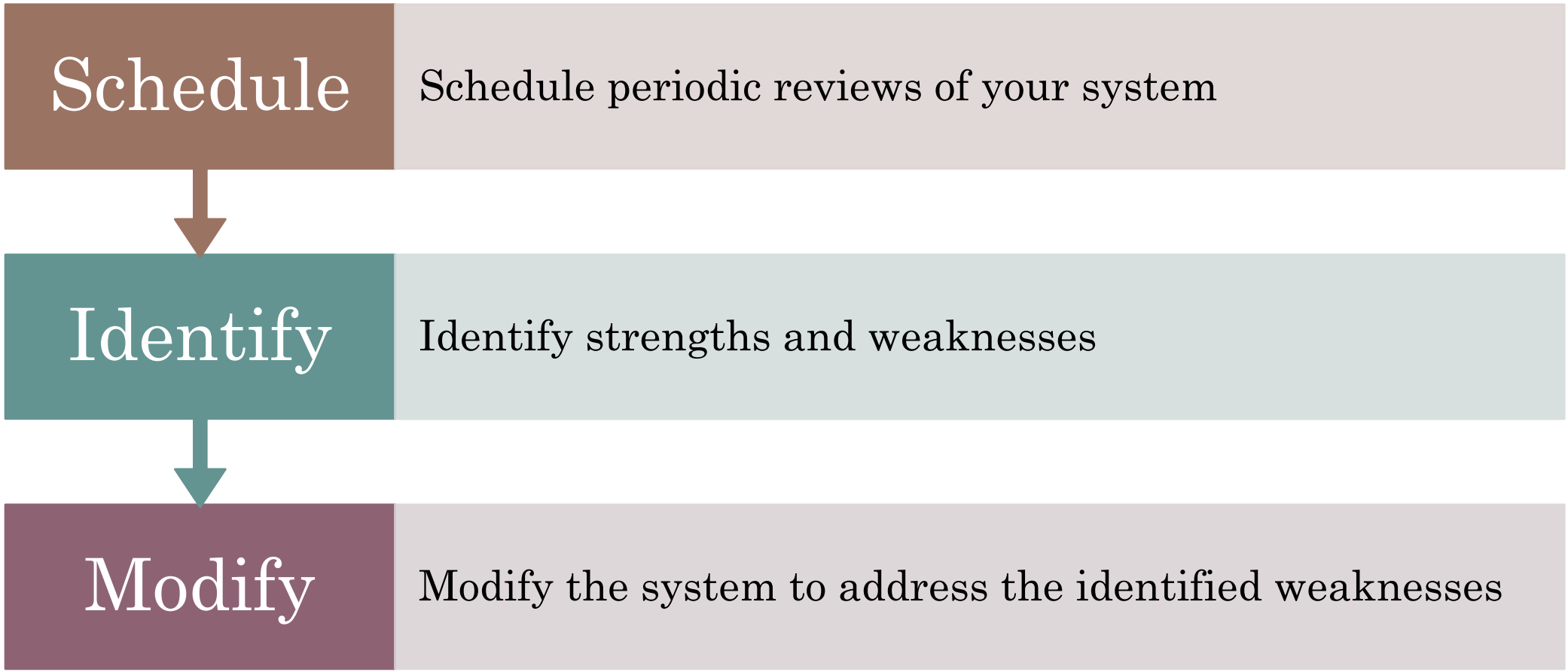
Examples of Feedback

- College of Pharmacy (Institutional Effectiveness report)
  - *Your report of Actions for Improvement for Goals 1 and 4 do not follow our guidelines for reporting. Please include who reviewed the results and state the actions to be taken as results of decisions made based on the review. Refer to your Goals 2 and 3 Actions for Improvement as examples of how this should be reported. Please avoid using any future tense phrases (will do, plan to do, etc.)*

Element 8: Plan for System Improvement

How will you review the effectiveness of your system, determine where improvements are needed, and implement the improvements?

Plan for Improvement



Let's Review the Elements

Element 1: Define the System and the Terms

Element 2: Establish the Institutional Framework

Element 3: Determine the System Inputs

Element 4: Determine the Plan and Report Structure

Element 5: Establish the System Processes

Element 6: Validity, Reliability, and Fairness

Element 7: Establish the System Outputs

Element 8: Plan for Improvement

Part 2: Unveiling the Mysteries of Reliability and Validity

To introduce and reinforce the concept of validity as it applies to higher education

To introduce and reinforce the concept of reliability as it applies in higher education

Validity

What it is, and how we examine it

- Validity refers to the *degree to which evidence and theory support the interpretations of the test scores for proposed uses of tests.* (p. 11)

Source:

- *American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (2014). Standards for educational and psychological testing. Washington, DC: AERA*

The Importance of Validity

Validity is, therefore, *the most fundamental consideration in developing tests and assessments.*

The process of validation involves accumulating relevant evidence to provide a sound scientific basis for the proposed score and assessment results interpretations. (p. 11)

Source:
AERA, APA, & NCME. (2014). Standards for educational and psychological testing. Washington, DC: AERA

How Do We Examine Validity in Higher Education?

Most often this is qualitative; colleagues are a good resource

Review the evidence

Common sources of validity evidence

- Test Content
- Construct (the idea or theory that supports the assessment)
- The validity coefficient

Important distinction

It is not the test or assessment itself that is validated, but the *inferences one makes from the measure based on the context of its use.*

Therefore it is not appropriate to refer to the 'validity of the test'; instead, we refer to the validity of the interpretation of the results for the test's intended purpose.

Validation Process

How do we establish that the interpretation and use of results of our assessments are valid for their intended purpose?

Validity Rationale: Building an *Interpretation and Use Argument*

When developing a validity rationale, it's important to state that the assessment covers relevant knowledge and skills (content validity).

However, validity rationales should also include support for the degree to which *inferences* you make from your results *match your assessment's intended purpose*

Building an Interpretation and Use Argument: **Purpose**

Describe the purpose of your assessment

What is its relationship to your curriculum?

Why this assessment is important to your program?

Building an Interpretation and Use Argument: **Content**

Does the measure adequately cover the content and skills students are expected to know and demonstrate?

Does the assessment measure what you've covered?

Building an Interpretation and Use Argument: **Context**

Is the assessment implementation strategy consistent with the purpose of the assessment?

Is the way the assessment is implemented going to allow students to yield their best performance?

Building an Interpretation and Use Argument: **Rubric Achievement Levels**

How do you know that the levels of achievement you developed for the rubric are appropriate for this assessment?

Do they adequately cover the potential responses you might hear?

Could any performance fall outside or in between your rubric levels?

Building an Interpretation and Use Argument: **Inferences**

How do you know that a rubric level matches your expectations for that level of performance?

How do you know that your use of test scores appropriately reflects the test's intended purpose?

Does your assignment of an achievement level or score *infer* that student has met the criterion at that level?

Review of the Validation Process – *The Interpretation and Use* Argument

- Establish the purpose of your test/assessment
- Develop content evidence
- Align the test/assessment context with its purpose
- Align the achievement levels and/or scores with their intended uses
- Establish the degree to which your assessment results infer the student's attainment of your established levels of achievement

Small Group Discussions

How would you establish validity in this situation?

Validity – Interpretation and Use Argument

- Your Biology faculty have established student learning outcomes for their program. Their content outcome is that students will:

Identify, describe and explain the basic terminology, concepts, methodologies and theories used within the biological sciences.

- They have selected the *Educational Testing Service (ETS) Biology Major Field Test* to measure this content outcome.
- This is a 3rd party exam used as a measure of a programmatic student learning outcome.
- What would be your interpretation and use argument for using this Field Test instead of a faculty-developed exam?

Reliability

What it is, and how we examine it

Reliability defined

- The general notion of reliability/precision is defined in terms of *consistency over replications* of the testing procedure. Reliability/precision is high if the scores/results for each person are consistent over replications of the testing procedure and is low if the scores are not consistent over replications. (p. 35, emphasis added)
- **Important:** *Replications* are defined to reflect a particular interpretation and use of test scores/assessment results.

Source:

- *AERA, APA, & NCME. (2014). Standards for educational and psychological testing. Washington, DC: AERA*

Basic Concepts Central to Establishing Reliability: **Correlation**

- **Correlation coefficient –** a statistical estimate of the strength and direction of the relationship between two continuous variables
- For every observed change in one variable, there is a related observed change in the other – they vary together
- These can be positive or negative, and the values range from -1 to +1
- We consider a test to be reliable if the coefficient is .70 or higher

Basic Concepts Central to Establishing Reliability: **True Score Theory**

- **True Score theory –** every person's **observed score** is a combination of the person's **true score** and some **measurement error**
- **True Score** formula:

Observed Score (X) = True Score (T) + Measurement Error (E)

- **True scores** are conceptualized as those obtained after *infinite* replications of the test/assessment
- We define the reliability coefficient as the **correlation** between **true scores** and **observed scores**
- Since we can never know a person's **true score**, we *approximate reliability* with a reliability coefficient – the **correlation between observed scores across repetitions**
- **Internal consistency** is typically a measure based on the *correlations* between different items on the same test

Basic Concepts Central to Establishing Reliability: **Types of Reliability Estimates**

- This number is estimated in several ways
- Parallel forms – the Pearson *product-moment correlation coefficient*
- Test-retest reliability – correlate scores on two administrations of the same test
- Split-halves reliability- Flanagan's Formula
- Kuder-Richardson Formulae 20 and 21 (KR_{20} and KR_{21}) – for tests with dichotomous items

A Common Approach to Estimating Reliability in Instructor-Made Tests: **Split-Half Reliability**

- First, split the test questions evenly into two parts – part a and part b, and calculate the subscores for each student on each part. Then, calculate the variances of subscores on the two parts and then the variance of the total scores. Enter these numbers into this formula and calculate:

$$r = 2\left(1 - \frac{S_a^2 + S_b^2}{S^2}\right)$$

- Key:
  - $S_a{}^2$ is the variance of part a
  - $S_b{}^2$ is the variance of part b
  - S^2 is the variance of total scores

Our example – History quiz results (max score = 10)

Name	*Quiz part a*	*Quiz part b*	*Total Score*
Joe	5	5	10
Mary	3	2	5
Cheryl	5	4	9
Charlene	4	4	8
Chris	2	3	5
Brian	3	4	7
LaTerrance	5	4	9
Kim	4	3	7
Robbie	3	5	8
Anwar	5	3	8

Calculating the Variances

To try this, use the standard deviation online calculator, and click "sample." Enter the scores for part *a*, part *b*, and the *total*, one at a time. You will get three variances.

Here are the results:

Part *a* = 1.21

Part *b* = .9

Total = 2.71

Next, we place these figures into the formula.

Calculating the Split-Half reliability

- Flanagan's Formula:

$$r = 2\left(1 - \frac{S_a^2 + S_b^2}{S^2}\right)$$

- Our data:

$$r = 2\left(1 - \frac{1.21 + .9}{2.71}\right) = 2\left(1 - \frac{2.11}{2.71}\right)$$

$$= 2(1 - .78) = 2(.22) = \underline{\mathbf{.44}}$$

Interpreting the coefficient

The Split-Half reliability is .44

Reliability coefficients should be at .70 or higher for a test to be considered reliable

This quiz is not reliable. What should the instructor do?

Another Approach to Reliability: Inter-rater Agreement

For open-ended decisions (e.g., rubrics, qualifying exams), have two readers score and examine the percentage of agreement

If agreement is low (< 70%, or .70), discuss where the disagreements occur and adjust scoring

Agreement cannot be at cost of validity or interpretability

Small Group Discussions

How would you interpret reliability coefficients in these situations?

Reliability – *How would you interpret these coefficients?*

An instructor has given a test in a Calculus 1 class. The instructor used the split-half reliability formula to calculate a reliability coefficient of .85. What does this tell you? What would you advise this instructor?

Two raters review a set of student research papers. The percentage of agreement is 55%, or .55. What would you advise the instructor?

Part 3: Examples from the Faculty

What do these tell you?

What, if anything, would you advise these faculty to do?

Example 1: Graduate Degree PhD in English

This is the 2017-18 Assessment Data Report of Program Goals (PGs) and Student Learning Outcomes (SLOs)

Use the Data Reporting Guide – have these faculty followed the guidelines?

What might you advise these faculty?

Example 2: Professional Degree Doctor of Medicine (MD)

This is the 2017-18 Assessment Data Report of Program Goals (PGs) and Student Learning Outcomes (SLOs)

Use the Data Reporting Guide – have these faculty followed the guidelines?

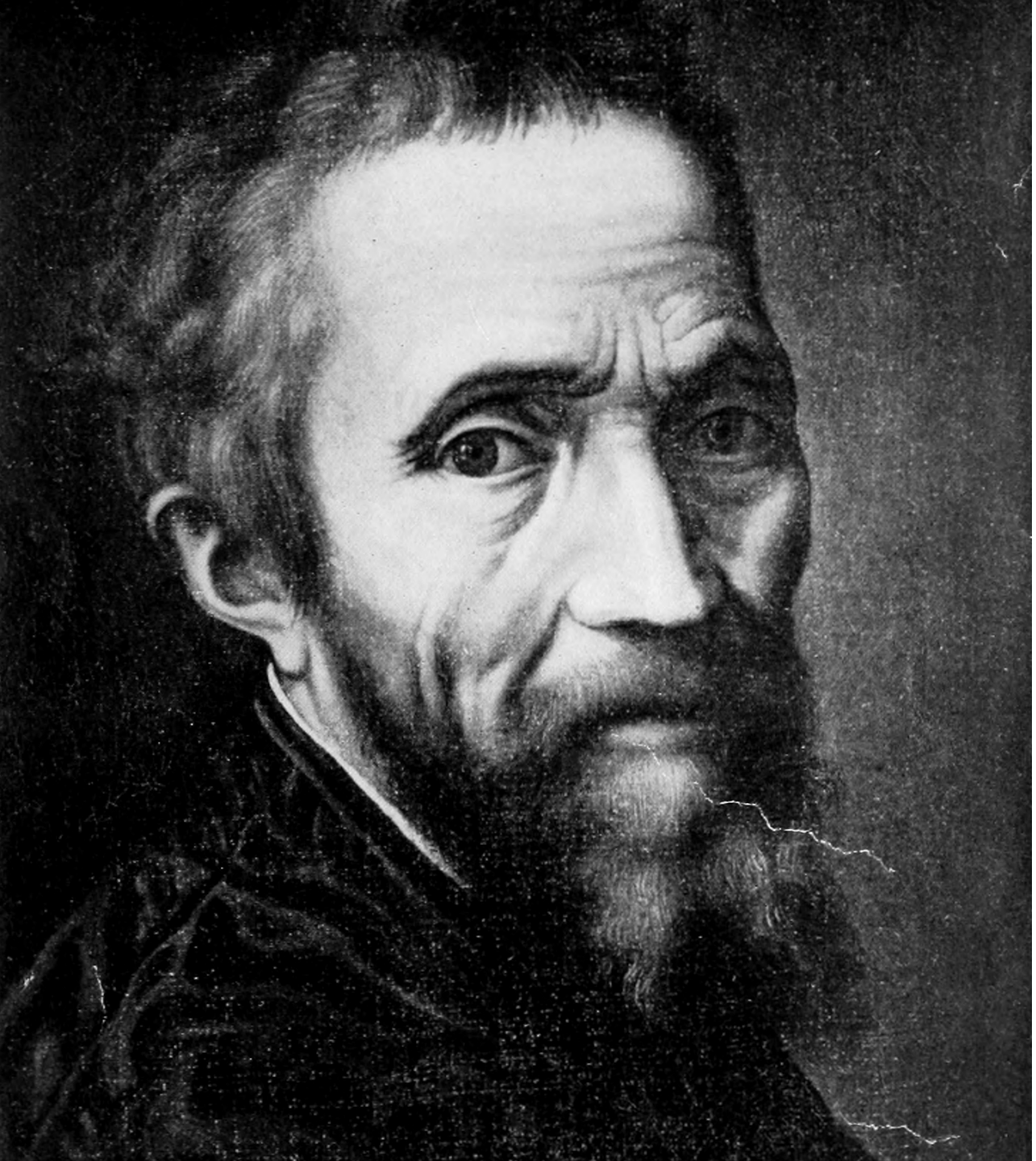
What might you advise these faculty?

Example 3: Undergraduate Degree Bachelor of Science in Journalism (BSJ)

This is the 2017-18 Assessment Data Report of Program Goals (PGs) and Student Learning Outcomes (SLOs)

Use the Data Reporting Guide – have these faculty followed the guidelines?

What might you advise these faculty?



A Closing Thought

The greater danger for most of us lies not in setting our aim too high and falling short; but in setting our aim too low and achieving our mark.

-Michelangelo

Source: https://www.michelangelo.org/michelangelo-quotes.jsp

THANK YOU!

Timothy S. Brophy, Ph.D.
Professor and Director, Institutional Assessment
tbrophy@aa.ufl.edu
+1-352-273-4476

Reference:

Brophy, T. S. (2017). The University of Florida Assessment System. In D. Miller & T. Cumming, (Eds.), *Enhancing assessment in higher education: Putting psychometrics to work* (pp. 184-202). Sterling, VA: Stylus.