SUNY OLD WESTBURY
OWN YOUR FUTURE

SUNY OLD WESTBURY
OWN YOUR FUTURE

PURSUING FAIRNESS AND TRUSTWORTHINESS IN ASSESSMENT DATA

JOSHUA FYMAN, M.S.
STEPHANIE SCHNEIDER, Ph.D.
SUNY OLD WESTBURY

SUNY OLD WESTBURY
OWN YOUR FUTURE

SUNY OLD WESTBURY
OWN YOUR FUTURE

(from 2020 Guide to Accreditation, AAQEP)

Fairness—It is vitally important that measures be equitable in representing performance of all stakeholders—including applicants, candidates, completers, and partners.

Providers must investigate evidence that the meaning of results differs across groups and consider that characteristics irrelevant to what is being measured or assessed may lead to differential outcomes.

Issues to consider related to fairness are the possible introduction of bias in assessment content or processes and other factors that might contribute to disparate access or outcomes for different groups.

SUNY OLD WESTBURY
OWN YOUR FUTURE

PURSUING FAIRNESS WITH PSYCHOMETRICS

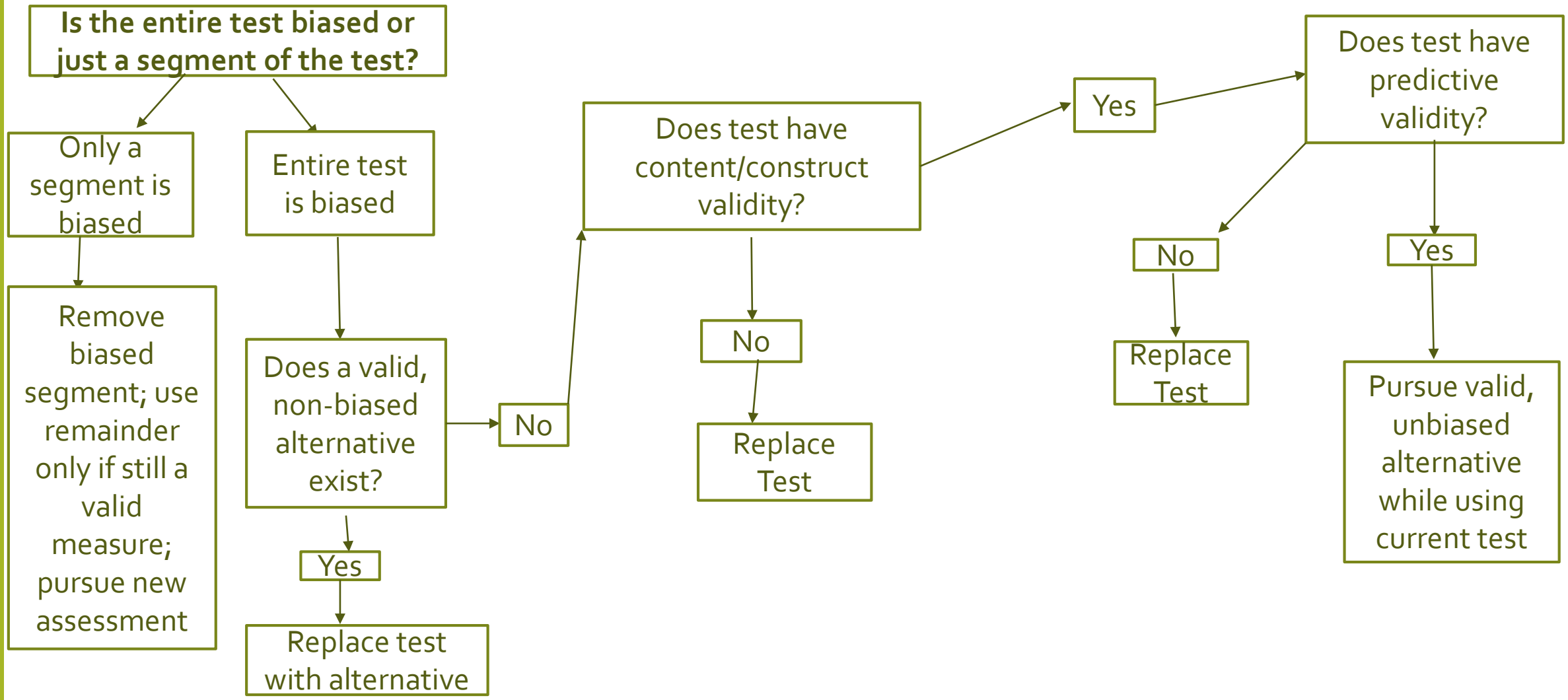
1. A FAIR ASSESSMENT MUST BE A VALID ASSESSMENT
    1. Content Validity – Does the assessment capture the entire domain being tested for? Does it measure things that are not part of the domain?

    2. Construct Validity – Does the assessment measure the construct it purports to measure?
    3. Predictive Validity –
        1. Can performance in this measure be linked to any future outcomes?
        2. If measure is not linked to any future outcomes, can its utility be explained?
    4. Group Differences-
        1. Is there a significant difference in performance on the measure between groups? (Gender, Race, Ethnicity, SES, etc.)

If Bias Is Found in Assessment Results…

SUNY OLD WESTBURY
OWN YOUR FUTURE

DECISION TREE



THE FAIRNESS STANDARDS APPLIED

SUNY OLD WESTBURY
OWN YOUR FUTURE

STANDARD	EXAMPLE
7.1 - When bias is discovered, validity testing must be examined.	7.1 – Comparison of means (t-test, ANOVA, etc.) discover significant difference between subgroups.
7.2 - When construct-irrelevant items or sections are identified, only sections of the exam that should be used, if any, are ones that demonstrate no disparate impact between subgroups.	7.2 – In otherwise valid test, disregard irrelevant section that favors male test-takers over females. Preferably, use other, valid test.
7.3 - When bias is discovered between subgroups of gender, race, ethnicity, disability, culture, or language, a team should launch a qualitative investigation of the potential sources of bias in the test design, content, and format and move to correct them.	7.3 – One subgroup performed significantly more poorly than all others on essay exam. Team reexamines the exam for any content, language, or method of administration that could be causing a problem for this subgroup's performance.
7.4 - Test developers should aim to identify language, symbols, words, phrases, and content that could cause unfair difficulty or disadvantage to subgroups of gender, race, ethnicity, disability, culture, or language.	7.4 - Team discovers that essay question relies on cultural references that a particular subgroup may be significantly less likely to be familiar with.
7.5 - A test taker's score should not be automatically considered a reflection of that examinee's ability without first considering other circumstances that may have affected the test taker's performance at that time.	7.5 – In considering a student's final exam, a reviewer should consider the fact that the weakest sections on the exam covered the month that he was missing classes due to family health concerns.
7.6 - When a criterion is predicted differently across subgroups, subgroup should be included as a moderating variable.	7.6 – A particular exam may needed to be weighted for race, sex, etc.

THE FAIRNESS STANDARDS APPLIED

SUNY OLD WESTBURY
OWN YOUR FUTURE

STANDARD	EXAMPLE
7.7 - When linguistic skills are not part of the construct of interest, language should be kept to a minimum in the test.	7.7 – Calculus exams should focus as much as possible on the calculations and not introduce too much language that can present difficulty for an ENL student on a construct they are not being tested on.
7.8 - When presenting disaggregated data across subgroups, it should be stated that scores may not be comparable across groups.	7.8 – If an exam had a 12% bias against non-white students, that fact should be mentioned when presenting the exam data.
7.9 - When test results are used for policy purposes, policy makers must be made aware of why scores may vary across subgroups.	7.9 – When selection test results are being shared to decide on a new cut score, it should be noted why certain subgroups are performing lower and what systemic problems may be contributing to that and thus warrant a reevaluation of the cut score.
7.10 - When the test results are being used to make policy that affects life outcomes for subgroups, policy makers must ensure that there is not another valid test for the construct that has less bias.	7.10 – If a selection exam is yielding different results for different subgroups, an alternative selection exam, provided it is valid, should be sought.
7.11 - When tests of a construct are of approximately equal validity, opt for the one with less bias.	7.11 – Two pre-Student Teaching Observation instruments have equal validity for predicting student teaching performance. One has significant difference of subgroup means while the latter does not. Opt for the latter.
7.12 - The testing and assessment phase should be conducted so that all subgroups experience equal treatment and equal opportunity throughout the process.	7.12 – The prep session for an exam were all scheduled for a day of the week that is a holy day for members of a particular religion. Every effort must be made to select days and times that do not discriminate in that manner.

Trustworthiness

- Measure used for qualitative data to examine four aspects:
  - Credibility
  - Transferability
  - Confirmability
  - Dependability

Credibility

- Truth Value
- What is the link to research?
- Like Construct Validity in Quantitative Research
- To Assess- member checks, triangulation, peer briefing

Transferability

- Applicability
- Will it work again?
- Like External Validity in Quantitative Research
- To Assess- thick description, purposive sampling

Dependability

- Consistency
- Can it be repeated?
- Like Reliability in Quantitative Research
- To Assess- create document trail and analysis, triangulation

Confirmability

- Neutrality
- Is this study neutral?
- To Assess- triangulation, reflective narrative