NACE Competency Assessment Tool
Info Sheet With Empirical Results
For Content Validity, Usability, Reliability, and Discriminant Validity

Content Validity
- **Asks:** Is the content in the instrument the appropriate content to include? Does the instrument cover the depth and breadth of the construct?
    1. Do the dimensions reflect the definition?
    2. Do the dimensions cover the range of the competency?
    3. Do the performance descriptors cover the range of each dimension?
    4. How essential is it to include each dimension?
- **Testing Population**: 373 practitioners over two rounds of data collection
- **Results:**
    o Across all four of these metrics, every assessment scored higher than 80% (minimum level to demonstrate content validity) and about half were more than 90%.

Usability
- **Asks:** Is the instrument usable? Specifically:
    1. Is the tool easy to use?
    2. Is the language student-friendly?
    3. Are the dimension titles written clearly and easy to understand?
    4. Are the performance descriptors written clearly and easy to understand?
    5. Is the level of detail appropriate?
- **Testing Population:** 104 students and 373 practitioners over two rounds of data collection.
- **Results:**
    o All assessments scored higher than 80% on questions 1-4.
    o The assessments scored an average of 48.3 on Question 5, where a score of 0 was "too little," 100 was "too much," and 50 was "just right."

Reliability
- **Asks:** How stable is the score?
    o Reliability measures the extent to which experts agree on the ratings, and high levels of agreement mean that experts view the performance similarly.
- **Testing Population:** 150 practitioners over two rounds of data collection.
- **Results:**
    o NACE used three metrics of reliability to assess each rubric:
        ▪ Simple percent agreement, which calculates the percent of time experts agreed (from 0% to 100%);
        ▪ Fleiss' Kappa, which is a chance-corrected measure of agreement because some agreement could happen by random chance (from -1 to +1); and
        ▪ The intra-class correlation (ICC), which is another standard way to assess expert agreement (from -1 to +1).

- o Interpretation:
  - ▪ Agreement: More than 80% is considered good, but it is also understood that this level of agreement should be interpreted in light of what is being measured. For example, on a simple two-point mathematics item, agreement should approach 100%. In contrast, when scoring a complex six-point writing constructed-response item, an agreement of 60% would be considered an acceptable result (NAEP, N.D.).
    - • Given that career readiness and these eight competencies are more complex constructs, lower levels of agreement are acceptable.
  - o Across these metrics, the rubrics generally demonstrated acceptable reliability. For example, all ICCs were higher than .80 and half were more than .90, which reflects good to excellent reliability. Percent agreement ranged from 59.2% to 84.2%, which reflects acceptable to strong reliability, and Fleiss's Kappa ranged from .40 to .70, which reflects moderate to substantial reliability. Please see the forthcoming technical report, which details all of these statistics.
  - o Therefore, we can say with confidence that the instruments are *reliable*.

Discriminant Validity
- • **Asks:** Is the instrument sensitive enough to distinguish (discriminate) between higher- and lower-skilled performances?
  - o If mean differences between the higher- and lower-skilled performances are statistically significant, then the instrument is deemed to demonstrate discriminant validity.
- • **Testing Population:** 150 practitioners over two rounds of data collection.
- • **Results:** All eight assessments showed evidence of discriminant validity. Please see the forthcoming technical report, which will detail and contextualize these results.

Conclusion:
- • Given the above evidence and the methodology employed, **we can conclude that the eight assessment tools are valid and reliable.** In sum, they demonstrate evidence for content validity, usability, reliability, and discriminant validity. NACE will release technical reports in the near future with more detailed statistics and information on the methodologies used.

#####