Salisbury University HEIghten Critical Thinking & Reasoning Assessment Report, Fall 2019

This report, authored by SU office of University Analysis, Reporting & Assessment (UARA) staff and reviewed by the University Academic Assessment Committee (UAAC), discusses Critical Thinking & Reasoning assessment data collected during fall 2019 GULL Week sessions.

To request more information about the assessment, results, or additional analyses, please contact the Assessment Coordinator, Dr. Sarah Winger.

Executive Summary

Background and Findings

1. Faculty and UARA agreed that the ETS HEIghten Critical Thinking (H-CT) assessment is aligned with the General Education student learning outcome, Critical Thinking & Reasoning.
2. The H-CT instrument comprises 26 items with an overall scaled score as well as 2 scaled subscores, for Analytic and Synthetic skills. The item formats include: critical thinking sets, short arguments or informational passages, and sets that present conditions applicable to a fictional situation. The item types include: single-selection, multiple choice; multiple-selection, multiple choice; select-in-passage; inline choice; and composite items.
3. The results of our administration of the H-CT instrument supported its validity and reliability.
    a. H-CT scores demonstrated validity:
        i. Content Validity: instrument was designed based upon literature review, review of existing measures, market research and survey of both higher education leaders and employers, as well as expert review of items
        ii. Scale Analysis: supported using both exploratory and confirmatory factor analyses
        iii. Criterion and Construct Validity: supported by published group differences, particularly based on class level (i.e., freshmen vs. seniors) and positive correlations with SAT/ACT scores– also, the overall score and subscores on this instrument had moderate to large positive correlations with the SU students' related measures of SAT Verbal score range categories and SAT Math score range categories
    b. H-CT scores in published studies satisfactorily supported reliability for both individual-level reliability and institution-level total score reliability of the overall score and subscores
4. A limitation of this administration is that some students (juniors) are overrepresented in this sample compared to other class levels – due to sampling requirements of a longitudinal study. Otherwise, generally, the students that completed the H-CT instrument were somewhat representative of the overall and non-test-taker populations at SU.
5. The average SU H-CT overall scaled score (161.8) was below the average of the comparison group (163.3) as well as below the proficiency benchmark (162). 48.8% of SU H-CT test-takers had scores below the 162-level proficiency benchmark. Similarly, the average SU H-CT scaled

Analytic and Synthetic subscores (4.0 and 4.0, respectively), were both below the averages of the comparison group (4.4 and 4.5, respectively).

6. There was no significant difference between H-CT overall scaled score, Analytic scaled subscore, or Synthetic scaled subscore averages of transfer students and SU native, first-time students.
7. As SU and other institutions' students' class level (i.e., freshmen, sophomores, juniors, seniors) increased, so too did the average H-CT overall scaled score (Table 12) and scaled subscores (Appendix 5 - Table 2). SU students' average H-CT overall scaled score increased significantly by class level; freshmen's average score was significantly less than juniors', seniors', and unclassified, non-degree undergraduates' average scores; sophomores' average score was significantly less than juniors' and unclassified, non-degree undergraduates' average scores. See full report for scaled subscore results.
8. There was a significant difference between average H-CT overall scaled score by SU college/school (i.e., CHHS, Fulton, Henson, Perdue, and Seidel; based on students' primary major); Henson majors' average score was significantly greater than average scores of students majoring in CHHS, Fulton, Perdue, or Seidel; no other college or school comparisons were significantly different. See full report for scaled subscore results.
9. Although the sample was small, a matched longitudinal analysis supported a significant increase in average H-CT overall scaled score for SU students over time – although there was no significant difference between the fall 2017 and fall 2018 averages.

Suggested Action Items

1. The benchmarks with which SU students' Critical Thinking & Reasoning are compared should be evaluated by objective faculty and/or staff with expertise in the discipline or assessment of those skills.
2. Perform an area/course mapping of the current SU courses that align with the revised Critical Thinking & Reasoning student learning outcome.
3. Teaching faculty, General Education Steering Committee, and other relevant parties should consider whether the H-CT instrument is well aligned with revised (as of November 2018) General Education Critical Thinking & Reasoning student learning outcome. If the H-CT instrument is not aligned, then an alternative assessment that is aligned should be identified.
4. Consider results from the assessment to develop interventions or review and update curriculum to align with areas that need improvement.
5. Relevant stakeholders at SU should request further analyses of the H-CT data to address additional questions of interest that were not described here.
6. Determine a timeline to re-collect assessment data related to Critical Thinking & Reasoning, tentatively set for re-assessing using the H-CT in fall 2022 and then every 3 years.
7. Continue collaborative longitudinal study and share updates with SU.

Contents

Detailed Critical Thinking & Reasoning Report

Instrument

ETS HEIghten Outcomes Assessment Suite

The ETS HEIghten Outcomes Assessment Suite comprises "innovative, modular, computer-delivered assessment tool[s that enable] colleges and universities to measure the student learning outcomes that are essential for academic success" (About the HEIghten Outcomes Assessment Suite 2020). The capitalized HEI in "HEIghten" stands for Higher Education Institution, indicating that the HEIghten Outcomes Assessment Suite instruments align with common general education areas in Higher Education. The instruments are designed and aligned with national frameworks, for the respective instruments. The instrument reporting includes score/subscore benchmark comparisons versus similar institutions. The benchmark comparison values in this report are from the most recently available ETS institutional score reports for the particular HEIghten assessment addressed.

H-CT Instrument

The HEIghten Critical Thinking (H-CT) assessment is one of the five ETS HEIghten Outcomes Assessment Suite instruments. The H-CT assessment is an instrument which comprises 26 items, whose formats include: "critical thinking sets, short arguments or informational passages, and sets that present conditions applicable to a fictional situation" (ETS HEIghten Critical Thinking Assessment 2020). Within those item formats are the following item types: single-selection, multiple choice; multiple-selection, multiple choice; select-in-passage; inline choice (i.e., drop-down menu items); and composite items. There are also other follow-up items (e.g., demographics, reason for taking test, did you try your best on this, etc.) following the 26 items. See H-CT sample items in Appendix 1 and information about the instrument's alignment with SU's student learning goals, outcomes, and curricular area mapping in Table 1. Details about the instrument can be found at the ETS HEIghten Critical Thinking Assessment website (2020), the ETS HEIghten Critical Thinking Test at a Glance document (2015), and the Liu *et al.* (2014) ETS Research Report, "Assessing Critical Thinking in Higher Education: Current State and Directions for Next-Generation Assessment" that explains the operational definitions and assessment considerations for the development of this particular assessment.

There are several indices which are measured by the H-CT and are described below. The first index is the overall scaled score as well as 2 scaled subscores, for Analytic and Synthetic skills. Analytical skills involve: 1) analyzing argument structure; 2) evaluating argument structure; and 3) evaluating evidence and its use. Synthetic skills involve: 1) developing valid or sound arguments; as well as 2) demonstrating understanding of the implications or consequences of information and argumentation. Each skill is aligned to approximately 50% of the H-CT items. Also, although it is not measured as a separate scaled subscore, the Understanding Causation and Explanation skill is embedded as a third dimension of critical thinking – within the overall scaled score. That skill involves understanding, evaluating, and creating arguments that invoke causal claims or that offer explanations for collections of information.

The University Academic Assessment Committee, representing Faculty Senate and multiple departments and programs, and UARA staff agreed that the H-CT instrument is aligned with the General Education Critical Thinking & Reasoning student learning outcome (Table 1).

Table 1. The SU General Education student learning goal, outcome, and area mapping related to Critical Thinking & Reasoning.

Student Learning Goal	Student Learning Outcome	Area Mapping
Essential Competencies	**Critical Thinking & Reasoning:** Students will be able to comprehensively analyze evidence before they create, critique, or accept an opinion, conclusion, or determine a need for further investigation.	TBD*

Notes. Revised SU General Education student learning goals and outcomes were approved by Faculty Senate November 20, 2018. Asterisk (*) denotes that, at this time, there has not been an official area mapping of current courses to the revised SU General Education student learning goals and outcomes.

Related to Critical Thinking & Reasoning, results from this instrument can: provide a benchmark of student outcomes at SU; inform instructional efficacy and possible interventions; evaluate curricular strengths and weaknesses; and continuously improve student outcomes if we use this instrument for future GULL Week administrations.

Methodology and Sample

Longitudinal Study of Student Learning Outcomes in Multiple Dimensions

Since GULL Week in fall 2017, SU has collaborated with ETS and other Higher Education Institutions to assess student learning outcomes over time in a longitudinal study.

- **Outcomes & Assessments:** Specifically, there were three types of student learning outcomes being assessed as a part of this study: generic (i.e., Critical Thinking; assessed with H-CT), "noncognitive" (e.g., organization, self-efficacy; assessed with SuccessNavigator), and domain-specific (i.e., Business-related knowledge; assessed with the ETS Major Field Test in Business). See Appendix 2 and Appendix 3 for more information on the latter two assessments, which are not assessed as part of the SU General Education.
- **Proctored Testing Sessions:** The first two outcomes were assessed during GULL Week in "longitudinal study sessions," although only certain students had access to these sessions (see study cohorts below) – whereas other students had access to GULL Week sessions that included different assessments that were aligned to other General Education student learning outcomes (non-longitudinal study sessions). Otherwise, the third outcome, assessed with the Major Field Test in Business was administered in separate proctored sessions by either UARA staff (fall 2017) or Perdue School of Business (hereafter, "Perdue") staff or graduate students (subsequent semesters, both fall and spring). Prior to participation in this study, Perdue had used and required student participation in the Major Field Test in Business as part of their capstone course, Management 492, which students generally take during their final semester prior to graduation. Perdue uses these results to evaluate and report upon student learning outcomes for their programs, both for informing internal continuing improvement of the programs as well as external accreditation required reporting.
- **Cohorts:** In fall 2017 there were two cohorts included in this study: 1) freshmen (all majors) or 2) junior Business majors. Subsequently, in fall 2018, those two cohorts were re-assessed – as sophomores (all majors) or senior Business majors, respectively. Finally, in fall 2019, only the former cohort was re-assessed, as juniors (all majors).
    - ***Sampling Limitations:*** Some students may have participated (or given consent to be included in the study) in only one GULL Week collection sample during the study. This could have been related to issues with GULL Week registration, participant turn-out,

attrition from SU, etc. For example, one common issue with the GULL Week registration system is that students need to manually adjust their class level from GULL Week to GULL Week so that they can access and enroll in the correct session type (see above). If the students do not manually update their class level, then they cannot sign up for the longitudinal study sessions and instead can only sign up for non-longitudinal study sessions that include different assessments that were aligned to other General Education student learning outcomes. Similarly, new students that had not participated in the original fall 2017 study may have joined the study in their sophomore or junior year. It is also the case that some students accidentally participated in a longitudinal study session when they were not in the correct class level/cohort for the specific GULL Week year and therefore should not have been included in the study (e.g., showed up to the wrong testing room and couldn't reschedule for a different time, therefore the proctor allowed them to take the H-CT and SN to get GULL Week participation credit; issues with GULL Week registration).

Fall 2019 GULL Week

Data were collected from volunteer students at SU who self-selected and signed up to participate in various Gaining Understanding as a Lifelong Learner (GULL) Week testing sessions during a week in September, 2019. GULL Week sessions were open to the entire SU undergraduate student population. The assessments were administered in a proctored computer lab setting and lasted approximately one hour, of which ~45 minutes was dedicated to the H-CT administration and the remaining ~15 minutes was dedicated, depending upon which session type, to either the SuccessNavigator instrument (Appendix 3; Markle *et al*. 2013; Rikoon *et al*. 2014; Rikoon & Midkiff 2018; longitudinal study sessions) or the Student Opinion Scale (SOS) Survey (Appendix 4; Sundre & Thelk 2007; non-longitudinal study sessions). The SuccessNavigator has subskill scores which align with student characteristics across four skills: academic skills, commitment, self-management, and social support (Appendix 3). These SuccessNavigator subskill scores were analyzed to evaluate if there were any relationships between those scores and the H-CT scores, for those students that participated in both. The SOS Survey estimates the GULL Week participant's perceived importance of the assessment(s) and effort expended by the participant in completing the assessment(s) (i.e., H-CT).

Some faculty offered incentives (such as extra credit) to participating students, some mentioned GULL Week and encouraged students to participate, and some did not interact with students about GULL Week. The office of University Analysis, Reporting & Assessment (UARA) publicized GULL Week across campus via many avenues. Particularly, competitions between both College/Schools and Greek life groups were set up to improve participation.

In all, n = 3022 undergraduates participated in fall 2019 GULL Week and, of those, n = 1281 students (18 years or older) completed the H-CT with quality data (39.3% and 16.7% of total SU fall 2019 undergraduate enrollment (n = 7686), respectively). The H-CT cut-off determination for "quality data" for the analyses in this report was based upon the UAAC decision of a student self-report measure of effort, informed by an *ad hoc* UARA analysis of various quality control metrics. Therefore, any student that self-reported less effort was marked as "not quality data" and therefore not included in these analyses. For the H-CT test, this is based upon the ETS follow-up question "Did you try your best?" and 49 students (3.7% of the total H-CT test-takers that included both "quality data" and "not quality data") that responded "No" were marked as "not quality data" and were only included as H-CT non-test-takers for these analyses. Similarly, the SuccessNavigator cut-off determination for "quality data" is imposed by

ETS, since any student that had particular quality data flags (i.e., "Minimum minutes answering questions too low" or "Percent completed in test too low") had "invalid" instead of a raw score for the 10 subskill values in the SuccessNavigator exported data file.

Demographic analyses of the H-CT non-test-takers (n = 6405; 83.3%), including those who participated without providing quality data, were compared to the test-takers that completed H-CT with quality data to evaluate the extent to which the sample of test-takers was representative of the entire SU undergraduate population during fall 2019.

Further analyses within the test-takers were performed to evaluate the validity and reliability of the instrument administration at SU as well as to determine whether scores on the instrument varied by student characteristic(s), based upon available data in the Student Information System (GullNet). Some of the data may be missing for some demographic or student data variables for some students, therefore some of these total numbers maybe different in the tables and results. However, there is a known limitation in the sampling because during fall 2019 GULL Week, more than half of the H-CT test-takers were juniors (all majors) as part of the longitudinal study, whereas sampling from freshmen, sophomore, and senior students were less represented in this H-CT test-taker sample. The students with data for both the H-CT and the SuccessNavigator were analyzed to evaluate student scores on those subskills. The students with data for both H-CT and the SOS Survey were analyzed to evaluate student responses on those scales.

Additionally, H-CT junior test-takers that participated in fall 2019 as well as the previous longitudinal study administrations of the H-CT (fall 2017 and fall 2018), were matched and analyzed (n = 53) to evaluate the impact SU curricula have on students' critical thinking skills over time. In GullNet, class level of students is defined by 30-credit increments (e.g., freshman = 0 – 30; sophomore = 31 – 60, etc.). However, some students take more or less than the recommended 30 credits per academic year, come to SU with additional credits transferred, enroll in summer/winter sessions, etc. which can cause variations in their class level variable as compared to their number of years enrolled at SU. To ensure the matched, longitudinal cohort was as large and accurate as possible, students who should have been juniors (e.g., based upon original date enrolled in SU courses) but were classified as sophomores or seniors for any reason (mostly due to completing too few or too many credits in the previous year) were retained in the junior cohort. This class level adjustment was similarly performed for the fall 2017 version of this cohort (freshman cohort) and fall 2018 version of this cohort (sophomore cohort; for more details, see Salisbury University H-CT Reporting Documentation for reports' names and locations).

Results

Demographic Comparison of Test-takers vs. Non-test-takers

Except for the limitations due to the designs of the targeted students in the previously mentioned longitudinal study sampling, the demographics of the students who took the H-CT were similar to the non-test-takers, based upon z-test results of column comparisons (Tables 2-8; lack of significance annotations). The impact of the longitudinal study's sampling targeting junior-level students are noted and evident in Table 5. Otherwise, Asians (Table 2), females (Table 3), SU native, first time students (Table 4), as well as both Henson and Seidel (Table 6) were disproportionately high in the test-taker group and, in two cases of student success metrics (i.e., High School GPA and SU Cumulative GPA), the test-takers of the H-CT were significantly more successful than the non-test-takers (Table 8); although it

should be considered that another set of success metrics (i.e., SAT math and verbal scores) indicated the two groups were mostly comparable (Table 7).

Table 2. Student Race/Ethnicity Compared between the H-CT Test-takers, Non-test-takers and All SU Undergraduates

Race/Ethnicity	Test-taker	Non-test-taker	Total
African American	165 (12.9%)	930 (14.5%)	1095 (14.2%)
American Indian/ Alaska Native	12 (0.9%)	47 (0.7%)	59 (0.8%)
Asian	65 (5.1%)*	225 (3.5%)*	290 (3.8%)
Caucasian	897 (70.0%)	4418 (69.0%)	5315 (69.2%)
Hispanic	59 (4.6%)	298 (4.7%)	357 (4.6%)
Native Hawaiian/ Pacific Islander	1 (0.1%)	8 (0.1%)	9 (0.1%)
Non-resident Alien	16 (1.2%)	77 (1.2%)	93 (1.2%)
Two or more races	25 (2.0%)	162 (2.5%)	187 (2.4%)
Unknown/ Not specified	41 (3.2%)	240 (3.7%)	281 (3.7%)
Total	1281 (100.0%)	6405 (100.0%)	7686 (100.0%)

Notes. Cell values are counts with percentages reported parenthetically. Significant difference of participation categories between test-takers' and non-test-takers' proportions are indicated by an asterisk (*), $p \leq .05$.

Table 3. Student Gender Compared between the H-CT Test-takers, Non-test-takers and All SU Undergraduates

Gender (code)	Test-taker	Non-test-taker	Total
Male (1)	391 (30.5%)*	3004 (46.9%)*	3408 (44.2%)
Female (2)	882 (68.9%)*	3366 (52.6%)*	4248 (55.3%)
Unknown/ Not specified	8 (0.6%)	35 (0.5%)	43 (0.6%)
Total	1281 (100.0%)	6405 (100.0%)	7686 (100.0%)

Notes. Cell values are counts with percentages reported parenthetically. Significant difference of participation categories between test-takers' and non-test-takers' proportions are indicated by an asterisk (*), $p \leq .05$.

Table 4. Student Admit Type, to SU, Compared between the H-CT Test-takers, Non-test-takers and All SU Undergraduates

SU Admit Type (code)	**Test-taker**	**Non-test-taker**	**Total**
First time student (F)	925 (72.8%)*	4231 (68.4%)*	5156 (69.1%)
Transfer (T + U)	346 (27.2%)*	1956 (31.6%)*	2302 (30.9%)
Total	1271 (100.0%)	6187 (100.0%)	7458 (100.0%)

Notes. Cell values are counts with percentages reported parenthetically. Significant difference of participation categories between test-takers' and non-test-takers' proportions are indicated by an asterisk (*), $p \leq .05$.

Table 5. Student Undergraduate Class Level Compared between the H-CT Test-takers, Non-test-takers and All SU Undergraduates

Class Level (code)	**Test-taker**	**Non-test-taker**	**Total**
Freshmen (1)	199 (15.5%)*	1751 (27.3%)*	1950 (25.4%)
Sophomores (2)	315 (24.6%)*	1343 (21.0%)*	1658 (21.6%)
Juniors (3)	553 (43.2%)*	1313 (20.5%)*	1866 (24.3%)
Seniors (and +) (4)	174 (13.6%)*	1705 (26.6%)*	1879 (24.4%)
Unclassified non-degree undergrads (7)	40 (3.1%)*	293 (4.6%)*	333 (4.3%)
Total	1281 (100.0%)	6405 (100.0%)	7686 (100.0%)

Notes. Cell values are counts with percentages reported parenthetically. Significant difference of participation categories between test-takers' and non-test-takers' proportions are indicated by an asterisk (*), $p \leq .05$. The disproportionately high number of juniors is caused by the majority of juniors being included instead in a separate concurrent study during fall 2019 GULL Week.

Table 6. Student College/School Enrollment Compared between the H-CT Test-takers, Non-test-takers and All SU Undergraduates

College/School	**Test-taker**	**Non-test-taker**	**Total**
CHHS	335 (26.2%)*	1355 (21.2%)*	1690 (22.0%)
Fulton	262 (20.5%)*	1829 (28.6%)*	2091 (27.2%)
Henson	227 (17.7%)*	900 (14.1%)*	1127 (14.7%)
Perdue	260 (20.3%)	1339 (20.9%)	1599 (20.8%)
Seidel	157 (12.3%)*	486 (7.6%)*	643 (8.4%)
Undeclared	40 (3.1%)*	496 (7.7%)*	536 (7.0%)
Total	1281 (100.0%)	6405 (100.0%)	7686 (100.0%)

Notes. Cell values are counts with percentages reported parenthetically. Significant difference of participation categories between test-takers' and non-test-takers' proportions are indicated by an asterisk (*), $p \leq .05$.

Table 7. Student SAT Scores Compared between the H-CT Test-takers, Non-test-takers and All SU Undergraduates

SAT Score Range	SAT Math			SAT Verbal		
	Test-taker	Non-test-taker	Total	Test-taker	Non-test-taker	Total
< 500	193 (22.1%)	962 (23.1%)	1155 (22.9%)	172 (19.7%)*	954 (22.9%)*	1126 (22.3%)
500-599	434 (49.7%)	2196 (52.7%)	2630 (52.2%)	426 (48.7%)	2070 (49.7%)	2496 (49.5%)
600-699	226 (25.9%)*	917 (22.0%)*	1143 (22.7%)	262 (30.0%)*	1036 (24.9%)*	1298 (25.8%)
700-800	21 (2.4%)	91 (2.2%)	112 (2.2%)	14 (1.6%)	106 (2.5%)	120 (2.4%)
Total	874 (100.0%)	4166 (100.0%)	5040 (100.0%)	874 (100.0%)	4166 (100.0%)	5040 (100.0%)

Notes. Cell values are counts with percentages reported parenthetically. Within each SAT subject, significant difference of participation categories between test-takers' and non-test-takers' proportions are indicated by an asterisk (*), $p \leq .05$. The SAT score ranges were used so that both the student scores on the old and 2016 SAT versions could be included.

Table 8. Student GPA Scores Compared between H-CT Test-takers and Non-test-takers

Success Metric	Test-taker		Non-test-taker	
	n	Avg (SD)	n	Avg (SD)
High School GPA	282	3.67 (.44)**	1887	3.56 (.47)**
SU Cumulative GPA	1274	3.18 (.59)**	6268	2.95 (.72)**

Notes. Cell values are sample sizes (n) or averages with standard deviation reported parenthetically. Significant difference of participation categories between test-takers' and non-test-takers' average values are indicated by two asterisks (**), $p < .001$.

Validity and Reliability of the H-CT Administration at SU

The results of our administration of the 26-item H-CT supported its validity and reliability. Much of the validity of the H-CT was described in ETS-related publications (Liu *et al.* 2014, Liu *et al*. 2016, Swiggett 2017). Content validity was supported via the steps of literature review, review of existing measures, as well as expert review of items (Liu *et al.* 2014, Swiggett 2017). This latter step also included a standard setting method to identify students' proficiency in Critical Thinking based on ranges in the overall scaled score and scaled subscores (Table 9; Swiggett 2017). Additionally, ETS performed market research and surveyed higher education leaders and employers to develop the H-CT (Guangming Ling, *personal communication*). Furthermore, scale analysis was supported using both exploratory and confirmatory factor analyses (Liu *et al*. 2016). Similarly, both individual-level reliability (Cronbach's alpha; α) and institution-level total score reliability satisfactorily supported the reliability of the overall score and subscores (Liu *et al*. 2016).

Table 9. Performance level descriptions and score/subscore interpretations for the H-CT (ETS HEIghten Critical Thinking Assessment Performance Level Descriptions 2016; ETS HEIghten Critical Thinking Sample Institutional Score Report. 2016; ETS HEIghten Outcomes Assessment Suite Scores 2020)

Score/Subscore Name	SU Proficiency Level	ETS Proficiency Level	Score/Subscore Range	A typical student at this level…
overall scaled score	Proficient	Advanced	173 – 180	…**has demonstrated the ability to:** • extrapolate implications • describe the logic of complex arguments • understand subtle logical relationships between assertions/arguments and supporting information • identify needed evidence and implicit assumptions • identify possible alternative causes or explanations
		Proficient	162 – 172	…**has demonstrated the ability to:** • make inferential connections • follow the logic of an argument • understand logical relationships between assertions/arguments and supporting information • identify implicit assumptions and evidence that supports or undermines a claim • distinguish causation from correlation
	Need Improvement	Developing	150 – 161	**…may:** • make inferential connections between two explicitly related points • follow the logic of an explicitly structured argument • mistake evidence that is broadly related to a topic for evidence that is relevant to a specific assertion • identify evidence that directly supports or undermines a claim • have difficulty distinguishing causation from correlation
Analytic scaled subscore	n/a	n/a	1 – 10	Varies
Synthetic scaled subscore	n/a	n/a	1 – 10	Varies

Note. Although the H-CT documentation describes the Advanced and Proficient proficiency levels, SU will only evaluate whether students are proficient or not and the "SU Proficiency Level" information details that difference.

Criterion and construct validity were supported based upon performance differences between freshmen and seniors at higher education institutions, while controlling for SAT/ACT scores, where seniors scored significantly greater than freshmen on average (Liu *et al*. 2016). Also, the SAT/ACT scores were correlated with the overall H-CT score. Based on the SU H-CT student scores in fall 2019, criterion and

construct validity were also supported because students' overall scaled score on this instrument had a large positive correlation with the SU students' related measure of SAT Verbal score range categories, *r* = .508 (p < .01) and a moderate positive correlation with the SU students' related measure of SAT Math score range categories, *r* = .430 (p < .01). Similarly, the students' Analytic scaled subscore on this instrument had moderate positive correlations with the SU students' related measures of SAT Verbal score range categories, *r* = .427 (p < .01), and SAT Math score range categories, *r* = .366 (p < .01). Also, the students' Synthetic scaled subscore on this instrument had moderate positive correlations with the SU students' related measures of SAT Verbal score range categories, *r* = .450 (p < .01), and SAT Math score range categories, *r* = .370 (p < .01). The SAT score range categories were from 1 - 4 where: 1 = < 500; 2 = 500-599; 3 = 600-699; and 4 =700-800). Correlation coefficients ≥ .3 but less than .5 are evidence of medium effect sizes and those ≥ .5 are evidence of large effect sizes (Field 2013).

SU Student Scores on the H-CT

On average, the students who participated (n = 1281) had an overall scaled score of 161.8 (SD = 6.1) with a range of 150 to 178 on the H-CT instrument (Table 10). For the Analytic skill, the average scaled subscore of participants was 4.0 (SD = 2.0) with a range of 1.0 to 9.8. For the Synthetic skill, the average scaled subscore of participants was 4.0 (SD = 2.0) with a range of 1.0 to 10.0. The possible overall scaled score range is 150 – 180 and the possible scaled subscores ranges are 1 – 10 (Table 9). The SU average overall scaled score is less than that of the comparison group, 163.3 (SD = 7.1), which comprises 16,224 undergraduate students of different class levels across 66 Higher Education institutions (either 2-year or 4-year institutions). Also, the H-CT proficiency levels (Table 9) indicate that improvement is needed since the SU average overall scaled score of 161.8 is less than 162, which is the benchmark cut-off for proficiency. Individual analysis indicates that 48.8% of the H-CT test-takers (n = 625) have overall scaled scores less than 162 and therefore need improvement. Similarly, both the SU average Analytic scaled and Synthetic scaled subscores (both 4.0) are less than as those of the comparison group, 4.4 and 4.5, respectively.

Table 10. SU (white columns) and Comparison Group (gray columns) Students' Proficiency Levels on the Scores/Subscores of the H-CT

Score/ subscore	**Comparison Group** (n = 16,224)			**SU Fall 2019** (n = 1281)		
	Avg (SD) ***SU Proficiency Level***	**Percent of Students**		**Avg (SD)** ***SU Proficiency Level***	**Percent of Students**	
		Proficient	**Need Improvement**		**Proficient**	**Need Improvement**
overall scaled score	163.3 (7.1) *Proficient*	58%	42%	161.8 (6.1) *Need Improvement*	51.2%	48.8%
Analytic scaled subscore	4.4 (2.3) *n/a*	n/a	n/a	4.0 (2.0) *n/a*	n/a	n/a
Synthetic scaled subscore	4.5 (2.3) *n/a*	n/a	n/a	4.0 (2.0) *n/a*	n/a	n/a

Notes. The comparison group data (gray) is based on the HEIghten Outcomes Assessment Suite Guide to Score Interpretation (2020). SU proficiency levels are: Proficient = students with scores ranging from 162 – 180; Need Improvement = students with scores ranging from 150 – 161 (see Table 9 for more details). Highlighted values denote where the SU fall 2019 scaled score or subscore averages were less than those of the comparison group.

On average, SU native, first time students had similar H-CT overall scaled scores as compared to transfer students (Table 11). The difference, .5, was not significant t(1269) = 1.25, p > .05. Similarly, the differences between first time students' H-CT scaled subscores as compared to transfer students' were

not significant: Analytic scaled subscore [-.01; t(1269) = -.09, p > .05] and the Synthetic scaled subscore [.2; t(1269) = 1.8, p > .05].

Table 11. Student Admit Type, to SU, Average Overall Scaled Scores on the H-CT

SU Admit Type (code)	**n**	**Score**	**SD**	**Percent of Students**	
				Proficient	**Need Improvement**
First time student (F)	925	161.9	6.0	52.3%	47.7%
Transfer (T + U)	346	161.4	6.2	48.3%	51.7%

Based on previous findings discussed regarding H-CT criterion and construct validity (Liu *et al*. 2016) as students' class level (e.g., freshman, senior) increased, so too did the average score on the instrument. The same trend was supported for this SU administration of the H-CT (Table 12). Specifically, at SU, juniors, seniors, and unclassified students scored greater than freshmen and sophomores on the H-CT overall scaled score. The difference in average scores between groups, although significant, was small based on effect size value interpretation (F(4, 1276) = 7.2, p < .001, *r* = .15). Post hoc comparisons, via the Tukey HSD test, were used to identify which class levels' average scores were significantly different. Tests revealed significant pairwise differences between the average scores of freshmen as compared to juniors, seniors, and unclassified, non-degree undergraduates (p < .05) as well as sophomores as compared and juniors and unclassified, non-degree undergraduates (p < .05).

In general, as class level increased, so too did the students' average scaled subscores (Appendix 5 - Table 2), with small effect sizes for each relationship [Analytic scaled subscore, (F(4, 229.896) = 6.0, p < .001, *r* = .14); Synthetic scaled subscore (F(4, 231.221) = 5.4, p < .001, *r* = .13)]. Post hoc comparison tests revealed significant pairwise differences between the average Analytic scaled subscores of freshmen as compared to juniors (p = .001) and unclassified non-degree undergraduates (p < .01) and sophomores as compared to unclassified, non-degree undergraduates (p = .05) as well as between the average Synthetic scaled subscores of freshmen as compared to seniors (and +) and unclassified, non-degree undergraduates (p < .05) as well as sophomores as compared to juniors, seniors, and unclassified, non-degree undergraduates (p < .05). No other class level comparisons for the average subscale scores were significantly different.

Table 12. Student Undergraduate Class Level Average Overall Scaled Scores on the H-CT

Class Level (code)	**n**	**Score**	**SD**	**Percent of Students**	
				Proficient	**Need Improvement**
Freshmen (1)	199	160.4[a]*	5.7	39.7%	60.3%
Sophomores (2)	315	161.0[a]*	5.7	47.3%	52.7%
Juniors (3)	553	162.3[ab]*	6.2	54.6%	45.4%
Seniors (and +) (4)	174	162.5[ab]*	6.2	56.3%	43.7%
Unclassified, non-degree undergrads (7)	40	164.3[b]*	6.9	70.0%	30.0%

Notes. Subset groups' average scores are indicated by group letters [a] and [b]. Where a class level differs significantly compared to another class level is indicated by an asterisk (*), p ≤ .05. Interpret results for the Unclassified, non-degree undergraduates group cautiously, as samples in the proficiency variable contain groups with fewer than 30 students.

Student performance by SU College or School is listed in Table 13. There was a significant difference in the H-CT overall scaled score based on enrollment in College or School at SU, but the effect of difference in average scores between groups was small based on effect size value interpretation (F(5, 1275) = 9.4, p < .001, *r* = .19). Post hoc comparisons, via the Tukey HSD test, were used to identify which College or Schools' average scores were significantly different. Tests revealed significant pairwise differences

between the average score of students from Henson, which was significantly higher, as compared to the average scores of students from the College of Health and Human Services (CHHS), Fulton, Perdue and Seidel ($p < .001$). The average scores of undeclared students as well as those from CHHS, Fulton, Perdue do not significantly differ ($p > .05$).

Similar relationships were supported between student enrollment and the average scaled subscores (Appendix 5 - Table 3), with small effect sizes for each relationship [Analytic scaled subscore, ($F(5, 1275) = 7.1$, $p < .001$, $r = .19$); Synthetic scaled subscore ($F(5, 1275) = 7.6$, $p < .001$, $r = .17$)]. Post hoc comparison tests revealed significant pairwise differences between the average Analytic scaled subscores of students from Henson as compared to undeclared students or students from CHHS ($p < .01$) as well as compared to students from Fulton, Perdue, and Seidel ($p < .001$). Post hoc comparison tests also revealed significant pairwise differences between the average Synthetic scaled subscores of students from Henson as compared to students from Perdue ($p < .01$) as well as compared to students from CHHS, Fulton, and Seidel ($p < .001$). No other college or school enrollment comparisons for the average subscale scores were significantly different.

Table 13. Student College/School Enrollment Average Overall Scaled Scores on the H-CT

College/School	n	Score	SD	Percent of Students	
				Proficient	Need Improvement
CHHS	335	161.6[a]**	6.1	49.3%	50.7%
Fulton	262	160.9[a]**	5.9	46.2%	53.8%
Henson	227	164.1[b]**	5.9	66.1%	33.9%
Perdue	260	161.5[a]**	5.9	50.4%	49.6%
Seidel	157	160.7[a]**	5.8	43.9%	56.1%
Undeclared	40	161.5[a]	6.5	50.0%	50.0%

Note. Subset groups' average scores are indicated by group letters [a] and [b], where the group [a] category differs significantly compared to group [b] category is indicated by two asterisks (**), $p < .001$.

Although not presented here, student performance by primary major is available upon request to programs or Departments when at least 30 students in that major participated in this instrument's administration. These data can be used for informal review and improvement efforts, or for more formal program review and improvement efforts such as Academic Program Review required reporting related to assessment of program student learning outcomes aligned with this instrument, when applicable.

H-CT and SuccessNavigator Student Scores

Some of the H-CT test-takers, those in the longitudinal study sessions, also took the SuccessNavigator with quality data (n = 564; Table 14). The reliability and validity of the instrument were examined by Markle *et al*. (2013). SuccessNavigator is an instrument which includes 94 Likert-style items (6-point, strongly disagree to strongly agree) that align with 4 skills and 10 subskills (Table 14; Appendix 3; Markle *et al*. 2013). The skill and subskill scores are standardized and scaled to a mean of 100 and a standard deviation of 15. For each score, the qualitative categories' ranges are based on the ETS benchmark sample population where: Low = the bottom 25 percent; Moderate = middle 50 percent; and High = top 25 percent.

In general, students' SuccessNavigator skill and subskill scores are indicative of the Moderate category. These indicate that, on average, there is not a high need for interventions or that the students are not at high risk for attrition from the institution. Furthermore, the correlation analyses of the SuccessNavigator

skills and subskills with the H-CT overall scaled score as well as the two scaled subscores indicated that there were significant correlations (Table 14). The H-CT overall scaled score had statistically significant small positive correlations with the *Self-Management* skill score (r = .136, p = .001) and its subskill scores*: Sensitivity to Stress* (r = .093, p < .05)*, Academic Self-Efficacy* (r = .090, p < .05), and *Test Anxiety* (r = .148, p < .001), The *Test Anxiety* and *Sensitivity to Stress* subskill scores are reverse-scored so that higher scores refer to lower test anxiety and stress responses, respectively. These correlation results indicate that students' critical thinking abilities are positively correlated with their reactions to academic and daily stress (Appendix 3). Specifically, students with higher H-CT overall scaled scores also have high *Self-Management* skills, including low *Sensitivity to Stress*, high *Academic Self-Efficacy*, and low *Test Anxiety*, as compared to the students with lower H-CT overall scale scores. The H-CT Analytic subscore had similar results, with a small positive correlation with the *Self-Management* skill score (r = .092, p < .05) and the *Test Anxiety* subskill score (r = .117, p < .01). These correlation results indicate that students with higher H-CT Analytic scaled subscores also have high *Self-Management* skills, including low *Test Anxiety*, as compared to the students with lower H-CT Analytic scaled subscores. The H-CT Synthetic subscore results were also similar, with small positive correlations with the *Self-Management* skill score (r = .130, p < .01) and the *Sensitivity to Stress* (r = .096, p < .05) and *Test Anxiety* subskill scores (r = .134, p = .001). These correlation results indicate that students with higher H-CT Synthetic scaled subscores also have high *Self-Management* skills, including low *Sensitivity to Stress* and low *Test Anxiety*, as compared to the students with lower H-CT Synthetic scaled subscores.

Table 14. SuccessNavigator Skill (gray) and Subskill (white) Average Scores for the Students who also Participated in the H-CT (n = 564)

SKILL/Subskill Name	Score	SD	Qualitative Category	H-CT Correlations (*r*)		
				Overall	Analytic	Synthetic
ACADEMIC SKILLS	104.2	14.2	Moderate			
Organization	104.5	15.2	Moderate			
Meeting Class Expectations	102.9	13.4	Moderate			
COMMITMENT	97.8	15.4	Moderate			
Commitment to College Goals	96.1	15.8	Moderate			
Institutional Commitment	99.6	15.0	Moderate			
SELF-MANAGEMENT	92.1	14.1	Moderate	.136***	.092*	.130**
Sensitivity to Stress[Rev]	94.5	14.9	Moderate	.093*		.096*
Academic Self-Efficacy	97.2	15.6	Moderate	.090*		
Test Anxiety[Rev]	91.6	13.7	Moderate	.148***	.117**	.134***
SOCIAL SUPPORT	99.9	14.8	Moderate			
Connectedness	100.1	15.6	Moderate			
Institutional Support	99.0	15.0	Moderate			
Barriers to Success[Rev]	101.0	15.1	Moderate			

Notes. "Rev" denotes subskills that are reverse-scored, where higher scores for these subskills represent positive probabilities to success and lower instances of stress sensitivity, test anxiety, or barriers to success. Similarly, higher values in the other skill/subskill scores represent positive probabilities to success. The effect sizes (r) of the correlations between the H-CT score and subscores and the SuccessNavigator skills and subskills have significance indicated by an asterisk (*), p ≤ .05; two (**), p ≤ .01; or three (***), p ≤ .001. Non-significant correlation values (p > .05) are not shown.

H-CT and SOS Survey Student Responses

Some of the H-CT test-takers, those in the non-longitudinal study sessions, also took the SOS Survey (n = 518; Table 15; Appendix 4). We were able to evaluate the reliability of both subscales within the SOS Survey. The *Importance* subscale, which addresses the extent to which the student thought it was

important to do well on the H-CT, demonstrated reliability (α = .756). Similarly, the *Effort* subscale, which addresses the extent to which the student fully engaged in effortful behavior on the H-CT, demonstrated reliability (α = .771). The validity of the instrument is discussed in the SOS Survey Manual (Sundre & Thelk 2007). The 10 items, five in each subscale, are measured in a 1 to 5 scale, where 1 = Strongly Disagree, 2 = Disagree, 3 = Neutral, 4 = Agree, and 5 = Strongly Agree. There are four items that are negatively worded, and their scores were reverse coded prior to analysis.

In general, students selected "Agree" in their responses for both the *Importance* and *Effort* subscales. For *Importance,* this indicates that students thought that their scores on the H-CT instrument would affect them somewhat in either a positive or negative way. For *Effort*, it indicates that students put in a moderate effort towards completing the H-CT instrument. The two subscales had a moderate positive correlation with one another, r = .408 ($p < .001$; medium effect size). There was evidence that both subscales were also minimally positively correlated with the H-CT overall scaled score [*Importance*, r = .088 ($p < .05$, small effect size); *Effort*, r = .278 ($p < .001$; small effect size)], although for the *Importance* subscale the BCA 95% CI (-.002, .173) indicate there is no clear correlational relationship with the H-CT overall scaled score. The *Effort* subscale correlation with the H-CT overall scaled score seems to indicate that the students who self-reported exerting more effort on their performance on the test also scored slightly greater than those who did not self-report exerting as much effort on the test, although the effect size was small.

Table 15. Student Opinion Scale (SOS) Survey subscales' administrative results for the students who also participated in the H-CT instrument administration

SOS Subscale	Number of Items	Reliability (α)	n	Average Score (out of 25)	SD
Importance	5	.756	518	16.4	3.5
Effort	5	.771	518	19.3	3.0

Longitudinal SU Student Scores on H-CT Instrument

There were 53 students who took the H-CT instrument, with quality data, during the fall 2017, fall 2018, and fall 2019 GULL Week administrations. Initial analyses (frequencies, kurtosis, skewness, and Kolmogorov-Smirnov test) indicated that the samples were normally distributed. Therefore, the parametric, one-way, repeated measures ANOVA was performed. H-CT overall scaled scores were significantly higher for the fall 2019 administration (average = 164.0, SD = 6.1) as compared to the fall 2017 (average = 160.9, SD = 6.0) or fall 2018 (average = 160.7, SD = 4.9) administrations. Mauchly's test indicated that the assumption of sphericity was not violated, $\chi^2(2) = .82$, $p > .05$, therefore, the one-way, repeated measures ANOVA was an appropriate analysis. The one-way, repeated measures ANOVA results show that the H-CT overall scaled score was affected by year the test was administered (i.e., class level at SU as students were sampled in their freshman, sophomore, and junior years) in a linear relationship with a large effect size, ($F(2, 104) = 11.41$, $p < .001$, r = .504. Paired samples T tests were used to follow up this finding for the pairs of years in the study. It appeared that H-CT overall scaled scores did not significantly change from the start of the study (fall 2017; freshman year of the cohort sample) to the second year of the study (fall 2018; sophomore year of the cohort sample), since this difference, -.21, BCa 95% CI [-1.24, 1.63], was not significant $t(52) = .272$, $p > .05$. However, the H-CT overall scaled scores did change significantly from the start of the study (fall 2017; freshman year of the cohort sample) to the third year of the study (fall 2019; junior year of the cohort sample), since the difference -3.08, BCa 95% CI [-4.57, -1.45], was significant $t(52) = -4.209$, $p < .001$, r = .504 (large effect size). Similarly, the H-CT overall scaled scores changed significantly from the second year of the study

(fall 2018; sophomore year of the cohort sample) to the third year of the study (fall 2019; junior year of the cohort sample), since the difference, -3.28, BCa 95% CI [-4.75, -1.65], was significant t(52) = -4.035, p < .001, *r* = .488 (medium effect size).

Furthermore, we considered the change in SU Proficiency Level, based upon the H-CT overall scaled score benchmarks (Table 9), of the 53 individuals that took the H-CT instrument across the three years (Table 16). In fall 2017, 56.6% of students (30) were included in the Need Improvement category. Of those 30 students, 9 students increased their score to the Proficient category in fall 2018. However, of the 23 students who originally scored in the Proficient category range in fall 2017, 9 students declined to the Need Improvement category range.

From fall 2018 to 2019, 16 of 30 students improved their scores from Need Improvement to Proficient. However, of the 23 students who scored in the Proficient category range for fall 2018, 3 students declined to the Need Improvement category range.

Looking across three years from fall 2017 to fall 2019, 14 students increased their score from Need Improvement to Proficient in fall 2019. Almost all students (22 of 23) who originally scored as Proficient remained in the Proficient category range for fall 2019. Overall, 30 students in Fall 2017 and Fall 2018 were in the Need Improvement category range, compared to only 17 students in Fall 2019. This indicates not only a statistically significant difference between scores over the last three years, but a practical increase in students' critical thinking skill proficiency from fall 2017 to 2019.

Table 16. SU Proficiency Level changes over time of the individuals that took the H-CT instrument across three years (n = 53)

H-CT administration	Percent (Count) of Students		Year-to-Year Change	Year Compared Count of Students		
	Proficient	Need Improvement		Fall 2017	Fall 2018	Fall 2019
Fall 2017	43.4% (23)	56.6% (30)	Improved	--	9	14
			Stayed the same		35	38
			Declined		9	1
Fall 2018	43.4% (23)	56.6% (30)	Improved	--	--	16
			Stayed the same			34
			Declined			3
Fall 2019	67.9% (36)	32.1% (17)	Improved	--	--	--
			Stayed the same			
			Declined			

Notes. The left half of the table shows the percentages and counts of the 53 students based upon their SU Proficiency Level (Table 9) by the semester and year that the H-CT was administered. The right half of the table compares the semester and year in question from the first half of the table to subsequent semester and year in which the H-CT was re-administered to those same 53 students and shows the count breakdown of how those students changed from year-to-year, based upon their compared SU Proficiency Level change from year-to-year. "Improved" indicates the number of students that changed from Need Improvement to Proficient; "Stayed the same" indicates the sum of students that were in the Need Improvement category range and stayed in the Need Improvement category range combined with those that were in the Proficient category range and stayed in the Proficient category range; "Declined" indicates the number of students that changed from Proficient to Need Improvement.

Discussion

Based on the results presented here it seems that there is room for improvement in the Critical Thinking & Reasoning student learning outcome at SU. Several action items are suggested below towards this end.

1. To determine whether our students are meeting SU expectations for Critical Thinking & Reasoning, the benchmarks with which SU students' Critical Thinking & Reasoning are compared should be evaluated by objective faculty and/or staff with expertise in the discipline or assessment of it. For example, what percentage of students do we expect to be proficient?
2. Perform an area/course mapping of the current SU courses that align with the revised Critical Thinking & Reasoning student learning outcome.
3. Based on discussions and decisions related to Action Items #1-2, relevant parties such as faculty teaching courses aligned with this student learning outcome and the General Education Steering Committee should consider whether the H-CT instrument is aligned well with the revised (as of November 2018) SU General Education Critical Thinking & Reasoning student learning outcome. If it is not aligned, then an alternative assessment that is aligned should be identified.
4. Relevant stakeholders at SU should consider the results from the H-CT assessment to develop interventions or review and update curricula to align with areas that need improvement. In particular, and possibly in conjunction with Action Item #2, these data can be re-evaluated to help identify particular courses that students with high H-CT score/subscores have completed at SU to investigate potentially successful Critical Thinking & Reasoning-related interventions on campus. Successful projects at other institutions may be considered to guide instructional interventions at SU.
5. Relevant stakeholders at SU should request further analyses of the H-CT data to address additional questions of interest that were not described here (e.g., potential analyses for particular courses or programs).
6. Based on discussions and decisions related to Action Items #1-5, a timeline for re-assessment of the SU General Education Critical Thinking & Reasoning student learning outcome should be finalized. At this time, the H-CT is planned to be re-assessed in fall 2022 and every three years after. This will allow an analysis of whether there is change in student learning outcomes based upon either a change in assessment or instructional or curricular interventions.
7. UARA will continue to collaborate with ETS on the "Longitudinal Study of Student Learning Outcomes in Multiple Dimensions" and share any subsequent findings, updates, or reports with stakeholders at SU.

References

About the HEIghten Outcomes Assessment Suite. 2020. https://www.ets.org/heighten/about/ Accessed September 2020.

ETS HEIghten Critical Thinking Assessment. 2020. https://www.ets.org/heighten/about/critical_thinking/ Accessed September 2020.

ETS HEIghten Critical Thinking Assessment Performance Level Descriptions. 2016. https://www.ets.org/s/heighten/pdf/critical-thinking-performance-level-descriptions.pdf Accessed September 2020.

ETS HEIghten Critical Thinking Sample Institutional Score Report. 2016. https://www.ets.org/s/heighten/pdf/critical-thinking-sample-institutions-score-report.pdf Accessed September 2020.

ETS HEIghten Critical Thinking Sample Items. 2020. https://www.ets.org/s/heighten/pdf/critical-thinking-sample-questions.pdf Accessed September 2020.

ETS HEIghten Critical Thinking Test at a Glance. 2015. https://www.ets.org/s/heighten/pdf/heighten_critical_thinking_test_at_a_glance_acc.pdf Accessed September 2020.

ETS HEIghten Outcomes Assessment Suite Scores. 2020. https://www.ets.org/heighten/scores/ Accessed September 2020.

Field, A. P. 2013. Discovering statistics using IBM SPSS Statistics (4th ed.). Thousand Oaks, CA: Sage Publications, Inc., pp. 267.

HEIghten Outcomes Assessment Suite Guide to Score Interpretation. 2020. ETS.

Liu, O. L., Frankel, L., & Roohr, K. C. 2014. Assessing Critical Thinking in Higher Education: Current State and Directions for Next-Generation Assessment. ETS Research Report Series, 2014(1), 1–23. https://doi.org/10.1002/ets2.12009

Liu, O. L., Mao, L., Frankel, L., & Xu, J. 2016. Assessing critical thinking in higher education: The HEIghten™ approach and preliminary validity evidence. Assessment & Evaluation in Higher Education, 41(5), 677–694. https://doi.org/10.1080/02602938.2016.1168358

Markle, R., Olivera-Aguilar, M., Jackson, T., Noeth, R., & Robbins, S. 2013. Examining Evidence of Reliability, Validity, and Fairness for the SuccessNavigator™ Assessment. ETS Research Report Series, 2013(1), i–58. https://doi.org/10.1002/j.2333-8504.2013.tb02319.x

Rikoon, S., Liebtag, T., Olivera-Aguilar, M., Robbins, S., & Jackson, T. 2014. A Pilot Study of Holistic Assessment and Course Placement in Community College: Findings and Recommendations. https://www.ets.org/research/policy_research_reports/publications/report/2014/jtra

Rikoon, S. H., & Midkiff, B. 2018. Using the SuccessNavigator Assessment to Assess Change Over Time: A Longitudinal Measurement Invariance Study. ETS Research Report Series, 2018(1), 1-9. https://doi.org/10.1002/ets2.12216

Swiggett, W. D. 2017. Providing Threshold Score Recommendations for the First Three Tests of the HEIghten® Outcomes Assessment Suite: A Standard-Setting Study. ETS Research Memorandum, 50. https://www.ets.org/Media/Research/pdf/RM-17-06.pdf

Sundre, D. L., & Thelk, A. D. 2007. "The Student Opinion Scale (SOS): A Measure of Examinee Motivation: Test Manual." James Madison University, The Center for Assessment & Research Studies.

Salisbury University H-CT Reporting Documentation

The following related reporting documentation can be found at the General Education Outcome Assessment Report website:

1. Fall 2017 Critical Thinking & Reasoning Assessment Report
2. Fall 2017 Critical Thinking & Reasoning ONESHEET
3. Fall 2018 Critical Thinking & Reasoning Assessment Report
4. Fall 2018 Critical Thinking & Reasoning ONESHEET
5. Fall 2019 Critical Thinking & Reasoning ONESHEET

Appendices

Appendix 1. ETS HEIghten Critical Thinking Sample Items

Note: These following sample items and answer key are for reference only and are originally from the ETS HEIghten Critical Thinking Sample Items document (2020). They provide examples of skills measured, contexts covered and the difficulty of the questions.

Questions 1 - 2 are based on the material below.

1. Records indicate that William Shakespeare was baptized on April 26, 1564, and buried April 25, 1616, in Stratford-upon-Avon, England.
2. There is no evidence that William Shakespeare attended school, but had he done so, it would have been the local grammar school, and he would have left by age 14.
3. Documents show that by the early 1590s William Shakespeare was a managing partner of the Lord Chamberlain's Men, an acting company in London that built the Globe Theatre.
4. A total of 37 plays list Shakespeare as the author, including 13 that are set in Italy and several that make references to London politics.
5. There is no evidence that Shakespeare traveled outside of England.
6. In writings by others during Shakespeare's lifetime, Shakespeare was often referred to as a writer.
7. There is no manuscript of any play in William Shakespeare's own handwriting; only print versions of his plays exist.
8. No one questioned Shakespeare's authorship of the plays attributed to him during his lifetime or for centuries after his death.
9. Christopher Marlowe (1564–1593) was a brilliant poet and dramatist, educated at Cambridge University, who pioneered blank verse (unrhymed lines, almost always in the pattern of stressed syllables called "iambic pentameter") for dramatic plays.
10. Blank verse praised for its beauty appears frequently in the works attributed to Shakespeare.
11. Edward de Vere, 17th Earl of Oxford (1550–1604), whose aristocratic crest of arms depicted a lion shaking a spear, was trained in law, was a court poet, and visited Italy extensively.

Argument 1: (an abstract of an academic paper in a literary journal)

Abstract: "William Shakespeare of Stratford Could Not Have Written the So-called Shakespearean Plays"

We all know that there was a real person named William Shakespeare, who was born in Stratford in 1564, the son of a middle-class glove-maker, and who died in 1616. He was also a well-known actor and managing partner of an acting company in London in the 1590s. Beyond that, there is not a shred of evidence linking him to the 37 plays ascribed to him. How could an uneducated actor

from Stratford have such intimate knowledge of court politics, legal matters, royalty, and Italy (the setting of 13 plays including *Othello, Merchant of Venice*, and *All's Well That Ends Well*)? Clearly, the plays reflect a sophisticated intellect, a familiarity with London politics, and a deep understanding of Latin and Greek literature—all improbable for a mere actor who grew up in Stratford and who had at best a grammar-school education. Either Edward de Vere (who is known to have visited Italy and was a court favorite) or Christopher Marlowe (who was college educated and the pioneer of blank verse for dramatic plays) was the real author of these brilliant and nuanced plays.

Argument 2: The argument below is a rebuttal in the form of a letter to the editor, published in a subsequent issue of the journal that published Argument 1 above.

It is ludicrous to question Shakespeare's authorship of the plays. The argument presented in this journal smacks of elitism. Other arguments for that position rely on conspiracy theory and convoluted logic. There is a historical record of such a man who was connected to London theater and whose name was given as the author of the plays. No one questioned Shakespeare's authorship until hundreds of years after his death. Those who put forward names of the "real" author—over 60 such names have been suggested—have their own agendas, including the elitism already mentioned, or a preference for a particular alternative author. Circumstantial evidence or outrageous ideas such as that Marlowe faked his own death in 1593 and authored some of the plays afterward, or that the real author, for whatever reasons, wanted to keep his own identity hidden, are flimsy and do not hold up under serious scrutiny.

1. Given the information in the facts list, someone wishing to establish that Marlowe is most likely the author of the plays attributed to Shakespeare would be aided in that task if which of the following were found and determined to be authentic? Select <u>all</u> that apply.

□ 1. Comparisons of Marlowe's plays with Shakespeare's plays that show strong linguistic parallels and similar range of vocabulary
□ 2. Journal entries in Marlowe's handwriting that note plot elements of a Shakespearean play prior to its being performed
□ 3. Historical events that continue into the 1600s and parallel key plot elements in the plays

2. From the following facts excerpted from the list, select the <u>two</u> that <u>together</u> most help to support a claim central to Argument 1.

□ 2. There is no evidence that William Shakespeare attended school, but had he done so, it would have been the local grammar school, and he would have left by age 14.
□ 4. A total of 37 plays list Shakespeare as the author, including 13 that are set in Italy and several that make references to London politics.
□ 5. There is no evidence that Shakespeare traveled outside of England.
□ 7. There is no manuscript of any play in William Shakespeare's own handwriting; only print versions of his plays exist.
□ 9. Christopher Marlowe (1564–1593) was a brilliant poet and dramatist, educated at Cambridge University, who pioneered blank verse (unrhymed lines, almost always in the pattern of stressed syllables called "iambic pentameter") for dramatic plays.

3. The following is an exchange between two contributors to an online literary forum.

Kate: Ursula Seti's undated poem "Eucalyptus," which compares the eucalyptus tree's periodic shedding of its bark to various momentous events in her own life, could not have been written before 1960. Before that date, Seti had never left her native Alaska, where it is far too cold for most species of eucalyptus trees to grow. In 1960, however, she visited Australia, where eucalyptus trees are very common, so the poem must have been written during or after that visit.

Miriam: But Seti could certainly have known that eucalyptus trees periodically shed their bark without having personally observed that process, so she could have written the poem at any time during her career, which began well before 1960.

Which of the following most accurately characterizes Miriam's response to Kate?

(A) It shows that Kate's argument assumes the very point that it attempts to demonstrate.
(B) It draws an opposing conclusion from the evidence cited in Kate's argument.
(C) It refutes Kate's argument by rejecting one of its unstated assumptions.
(D) It calls into question one of the statements Kate makes to support her conclusion.

4. In Longport, a survey of residents showed that more of them had taken continuing education classes in literature than in the arts over the last twelve months. If so, some residents must have taken multiple arts classes, because an examination of enrollment figures showed that overall enrollment in continuing education arts classes was higher than overall enrollment in continuing education literature classes.

The reasoning in the passage depends on assuming which of the following?

(A) There was no substantial enrollment in arts classes by people who were not residents of Longport.
(B) There were no more literature classes than arts classes.
(C) Few, if any, residents took both an arts class and a literature class in the last twelve months.
(D) Most Longport residents took at least one arts class in the last twelve months.

Questions 5 - 6 are based on the information below.

In a benefit concert, seven solo performers—Harris, Jones, McIntyre, Nelson, Strapp, Trevino, and Williams—will each sing once only and one after another. The order in which the performers will sing is governed by the following conditions:

Harris must sing at some time before McIntyre sings.
Strapp must sing at some time before Jones sings.
Trevino must sing either immediately before or immediately after Nelson sings.
Williams must sing third.

5. If McIntyre is to sing immediately before Strapp sings, Trevino can sing
    (A) second
    (B) fourth
    (C) sixth
    (D) seventh

6. If McIntyre is to sing fourth, which of the following must be true?
    (A) Harris sings at some time before Strapp sings.
    (B) Jones sings at some time before Trevino sings.
    (C) Nelson sings at some time before McIntyre sings.
    (D) Strapp sings at some time before Williams sings.

Answer Key
1) 1, 2
2) 4, 5
3) C
4) A
5) A
6) D

Appendix 2. Information about Additional Assessments in the Longitudinal Study of Student Learning Outcomes in Multiple Dimensions: SuccessNavigator and the Major Field Test in Business

SuccessNavigator

- *Note: SuccessNavigator has been discontinued for institutional purchase as of September 2019 and the ETS website resources dedicated to it were removed in October 2020.*
- For information regarding the development and utility of the instrument, please see the related references above (i.e., Markle *et al*. 2013 for the former and Rikoon *et al*. 2014 as well as Rikoon & MidKiff 2018 for the latter)
- See more details regarding the instrument in Appendix 3

ETS Major Field Test for the Bachelor's Degree in Business (MFT-B)

- General Information about the ETS Major Field Tests (MFT):
  - About the MFT: https://www.ets.org/mft/about/
  - Design of the MFT: https://www.ets.org/mft/about/test_design/
  - MFT Research: https://www.ets.org/mft/about/research/
  - MFT FAQs: https://www.ets.org/mft/faq/
- Specific Information about the MFT-B:
  - About the MFT-B: https://www.ets.org/mft/about/content/bachelor_business - including the details that the assessment, "contains 120 multiple-choice questions designed to measure a student's subject knowledge and the ability to apply facts, concepts, theories and analytical methods. Some questions are grouped in sets and based on diagrams, charts and data tables. The questions represent a wide range of difficulty and cover depth and breadth in assessing students' achievement levels."
  - MFT-B Test Description document: https://www.ets.org/s/mft/pdf/mft_testdesc_business.pdf
  - MFT-B Sample Questions document: https://www.ets.org/s/mft/pdf/mft_samp_questions_business.pdf
  - 9 MFT-B Assessment Indicators (aka discipline-specific subscales) from the Departmental Summary of Assessment Indicators Sample Report document (https://www.ets.org/Media/Tests/MFT/pdf/MFT_sample_reports_2007/BusinessAISummary.pdf):
    1. Accounting
    2. Economics
    3. Management
    4. Quantitative Business Analysis
    5. Finance
    6. Marketing
    7. Legal and Social Environment
    8. Information Systems
    9. International Issues

Appendix 3. SuccessNavigator Construct Map and Question Details (modified from: Appendix from Markle *et al*. 2013 and Table 1 and Table 2 from Rikoon & Midkiff 2018)

Skill	Skill Definition	Subskill	Subskill Definition	Questions/ Subskill	Example Questions
Academic Skills	Tools and strategies for academic success	Organization	Strategies for organizing work and time	9	• I write a daily to-do list. • I use a calendar to plan my school day.
		Meeting Class Expectations	Doing what's expected to meet the requirements of courses including assignments and in-class behaviors	10	• I am on time for class. • I complete my assignments on time.
Commitment	Active pursuit toward an academic goal	Commitment to College Goals	Perceived value and determination to succeed in and complete college	10	• One of my life goals is to graduate college. • The benefit of a college education outweighs the cost.
		Institutional Commitment	Attachment to and positive evaluations of the school	8	• This is the right school for me. • I'm proud to say I attend this school.
Self-Management	Reactions to academic and daily stress	Sensitivity to Stress[Rev]	Tendency to feel frustrated, discouraged or upset when under pressure or burdened by demands	10	• I get stressed out easily when things don't go my way. • I am easily frustrated.
		Academic Self-Efficacy	Belief in one's ability to perform and achieve in an academic setting	9	• I'm confident that I will succeed in my courses this semester. • I can do well on tests if I apply myself.
		Test Anxiety[Rev]	General reactions to test-taking experiences, including negative thoughts and feelings (e.g., worry, dread)	9	• When I take a test, I think about what happens if I don't do well. • The night before a test, I feel troubled.
Social Support	Connecting with people and student resources for success	Connectedness	A general sense of belonging and engagement	7	• I feel connected to my peers. • People understand me.
		Institutional Support	Attitudes about and tendency to seek help from established resources	11	• If I don't understand something in class, I ask the instructor for help. • I know how to find out what's expected of me in classes.
		Barriers to Success[Rev]	Financial pressures, family responsibilities, conflicting work schedules, and limited institutional knowledge	11	• Family pressures make it hard for me to commit to school. • People support me going to college.

Notes. "Rev" denotes subskills that are reverse-scored. Higher scores for these subskills represent positive probabilities to success and lower instances of stress sensitivity, test anxiety, or barriers to success (e.g., the *Test Anxiety* and *Sensitivity to Stress* subskill scores are reverse-scored so that higher scores refer to lower test anxiety and stress responses, respectively).

Appendix 4. Student Opinion Scale (SOS) Survey (Sundre & Thelk 2007)

Item	Item Text	Subscale
1	Doing well on these tests was important to me.	Importance
2	I engaged in good effort throughout these tests.	Effort
3*	I am not curious about how I did on these tests.	Importance
4*	I am not concerned about the scores I receive on these tests.	Importance
5	These were important tests to me.	Importance
6	I gave my best effort on these tests.	Effort
7*	While taking these tests, I could have worked harder on them.	Effort
8	I would like to know how well I did on these tests.	Importance
9*	I did not give these tests my full attention while completing them.	Effort
10	While taking these tests, I was able to persist to completion of the tasks.	Effort

Note. Asterisk (*) denotes items that are reversed prior to scoring.

Appendix 5. Additional H-CT scaled subscore results by demographic groups

Appendix 5 - Table 1. Student Admit Type, to SU, Average Scaled Subscores on the H-CT

Scaled Subscore	**SU Admit Type** (code); sample size			
	First time student (F); n = 925		**Transfer** (T + U); n = 346	
	Score	**SD**	**Score**	**SD**
Analytic	3.9	2.0	4.0	2.0
Synthetic	4.1	1.9	3.9	2.0

Note. Neither scaled subscore was supported by significant differences of categories' average values, $p > .05$.

Appendix 5 - Table 2. Student Undergraduate Class Level Average Scaled Subscores on the H-CT

Scaled Subscore	**Class Level** (code); sample size									
	Freshmen (1); n = 199		**Sophomores** (2); n = 315		**Juniors** (3); n = 553		**Seniors (and +)** (4); n = 174		**Unclassified, non-degree undergrads** (7); n = 40	
	Score	**SD**	**Score**	**SD**	**Score**	**SD**	**Score**	**SD**	**Score**	**SD**
Analytic	3.5^{a*}	1.9	3.8^{a*}	2.0	4.1^{ab*}	2.0	4.0^{ab}	2.1	4.7^{b*}	2.5
Synthetic	3.7^{a*}	1.8	3.8^{a*}	1.8	4.2^{ab*}	2.0	4.3^{ab*}	2.0	4.7^{b*}	2.2

Notes. Subset groups' average scores are indicated by group letters [a] or [b]. By respective scaled subscore, where a class level differs significantly compared to another class level is indicated by an asterisk (*), $p \leq .05$, although one pairwise difference is significant at $p = .001$ [i.e., Analytic scaled subscore averages of Freshmen vs. Juniors].

Appendix 5 - Table 3. Student College/School Enrollment Average Scaled Subscores on the H-CT

Scaled Subscore	**College/School**; sample size											
	CHHS; n = 335		**Fulton**; n = 262		**Henson**; n = 227		**Perdue**; n = 260		**Seidel**; n = 157		**Undeclared**; n = 40	
	Score	**SD**	**Score**	**SD**	**Score**	**SD**	**Score**	**SD**	**Score**	**SD**	**Score**	**SD**
Analytic	4.0^{ab*}	2.0	3.8^{a*}	1.9	4.6^{b*}	2.0	3.8^{a*}	2.0	3.6^{a*}	2.0	3.6^{a*}	2.1
Synthetic	3.9^{a*}	2.0	3.7^{a*}	1.8	4.7^{b*}	2.0	4.0^{ab*}	1.9	3.8^{a*}	1.9	4.2^{ab}	1.9

Notes. Subset groups' average scores are indicated by group letters [a] or [b]. By respective scaled subscore, where a College/School differs significantly compared to another College/School is indicated by an asterisk (*), $p < .05$, although some pairwise differences are significant at $p < .001$ [e.g., Henson vs. some other College/School categories].