Psychometric Evaluation of the HAPS Anatomy & Physiology Comprehensive Exam

White paper presented by

Elizabeth Witt, PhD,
Witt Measurement Consulting
and the HAPS Testing Task Force

January 2017

Executive Summary

The HAPS A&P Comprehensive Exam is administered objectively and fairly under secure testing conditions. Based on expert psychometric analysis, the HAPS A&P Comprehensive Exam has been demonstrated to be reliable and valid for the purpose of measuring the HAPS learning outcomes. As a valid measure of the learning outcomes, the exam may be used to compare the performance of various groups of test takers. Scores are also sufficiently reliable to use in evaluating individual student performance relative to the HAPS learning outcomes.

The mission of the Human Anatomy and Physiology Society (HAPS) is to promote excellence in the teaching of anatomy and physiology. HAPS offers the A&P Comprehensive Examination as a measure of knowledge appropriately covered in two semester undergraduate courses in Anatomy and Physiology. This comprehensive exam is intended to assess the body of knowledge and is not tied to a specific curriculum. Rather, it is designed to reflect the HAPS learning outcomes. The purpose of the exam is to provide means for schools to compare their students' collective performance with the normalized data accumulated from the results of all students that have taken the same exam. Results may also be useful in evaluating the effectiveness of curriculum and instruction relative to the HAPS learning outcomes.

Exam Description

The exam is administered online under secure conditions. It consists of 100 high-quality multiple-choice questions with five answer options each. Questions are constructed to reflect both lower order (knowledge & comprehension) and higher order Bloom's taxonomy domains. Content is aligned with the HAPS learning outcomes, with module weights reflected in the number of questions as shown in Table 1.

Table 1. Exam Content Outline, Weighted by HAPS Learning Outcomes

Module		Weight
A	Body Plan & Organization	2%
B	Homeostasis	1%
C	Chemistry & Cell Biology	6%
D	Histology	3%
E	Integumentary System	2%
F	Skeletal System & Articulations	5%
G	Muscular System	6%
H	Nervous System	14%
I	Special Senses	4%
J	Endocrine system	6%
K	Cardiovascular System	14%
L	Lymphatic System and Immunity	4%
M	Respiratory System	7%
N	Digestive System	6%
O	Metabolism	3%
P	Urinary System	5%
Q	Fluid/Electrolytes & acid/base balance	7%
R	Reproductive System	5%
		100%

The HAPS learning outcomes for anatomy and physiology are described in detail in the HAPS website.

History

The HAPS Comprehensive Examination was originally established in June 1993 as a standardized assessment for evaluating course effectiveness. The exam was originally administered on paper, but the paper format and content was phased out due to security concerns. The paper exam was replaced with a new comprehensive exam (with new questions) developed by the HAPS Testing Committee, under the leadership of Eric Sun and Curtis DeFriez, past Co-Chairs of the HAPS Testing Committee. All questions were thoroughly reviewed, refined, and pilot tested. The exam was updated with new questions in 2010, and two forms were created. At that time, the exam administration was also changed to a secure, online delivery format hosted by ChiTester software.

The exam is continually reviewed and periodically updated by the HAPS testing task force to ensure that it remains relevant to the current knowledge base and a valid measure of the HAPS learning outcomes. The testing task force (i.e., a smaller subgrouping of the larger HAPS Testing Committee) consists of leaders in the field—including educators, item writers, and noted textbook authors in anatomy and physiology—all of whom have significant experience in testing as well. (See Appendix A for credentials of HAPS testing ask force members.) In addition to reviewing the content of questions and linkage to the learning outcomes, the testing task force makes use of item analyses and test summary statistics to monitor the quality of the exam.

Validity

Validity is defined as "the degree to which evidence and theory support the interpretations of test scores for the proposed uses of tests" (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014, p. 11). In more common terms, validity is the degree to which a test measures what it claims to measure and can appropriately be used for its intended purposes.

The primary source of evidence for the validity of the HAPS comprehensive examination is based on content. The exam is designed by subject matter experts (SMEs) to assess the HAPS learning outcomes and is carefully constructed according to the content specifications shown above. Each test question is explicitly aligned with one of the HAPS learning modules and categorized according to Bloom's taxonomy. All questions are written by experts in anatomy and physiology, who make use of standard guidelines on writing multiple-choice test questions to ensure consistency of format across questions. Question writers receive training in the art of item writing and use the guidelines developed by the National Board of Medical Examiners (NBME, 2002). The testing task force reviews all questions for content alignment, readability, fairness, and unnecessary jargon. In developing the exam, a pilot study was conducted, followed by an item analysis; any questions exhibiting potential problems were modified. Item analyses are conducted regularly, and questions are modified or removed if they do not perform well according to psychometric standards.

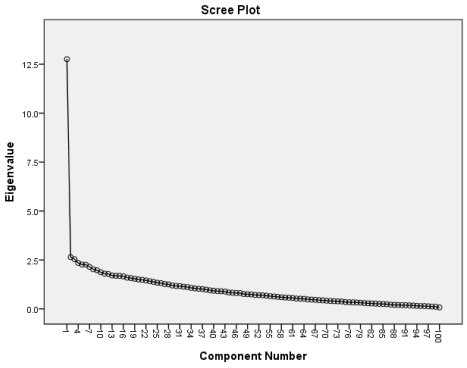
Evidence of validity based on internal structure is found in the relatively high levels of coefficient alpha and point-biserial correlations (see section on Item Analysis, below), indications that the questions included in the test are closely related to the overall construct being tested and discriminate well

between students with more knowledge and those with less knowledge. A principal components analysis also provides strong evidence that both the Red and Green forms of the examination primarily assess a single construct: knowledge of anatomy and physiology, or mastery of the HAPS learning outcomes. The scree plots shown here indicate that a single factor stands out. A scree plot displays the results of correlations among items as components or factors, with eigenvalues presented from largest to smallest. The larger the eigenvalue, the more variance is accounted for by the factor. As these plots show, one factor has a much larger eigenvalue than any other component. In fact, there is a steep drop from the first to the second component in terms of variance accounted for. The term "scree plot" is derived from its resemblance to scree, defined as an accumulation of loose stones lying at the base of a cliff or mountain. One might think of the components with low eigenvalues as having relatively little value, akin to the rubble at the base of a cliff.

Figure 1. Scree Plots Showing Relative "Value" of Factors for Red and Green Forms

Scree Plot – Red Form

Scree Plot – Green Form



Evidence of validity based on consequences is also accumulating as classrooms use the HAPS examination to measure learning gains. Students' mastery of the HAPS learning outcomes is demonstrably improved following instruction in anatomy and physiology. As an example, a study conducted at Emory University during the 2015-2016 academic year showed a mean increase of approximately 23 points among the 56 students who completed two semesters of coursework in the content area and took both the pre-test and the post-test.

Reliability

Reliability refers to the consistency or precision of test scores. The more reliable test scores are, the more likely an individual student would receive the same score (or a very similar score) if testing on a different day or using a different form of the test. It can be impractical, however, to obtain and correlate scores from a large group of students on two forms or two different days. Therefore, test

reliability is typically estimated via a statistic measuring internal consistency, most commonly Cronbach's coefficient alpha. Alpha is based on the correlations among test questions; the higher the value, the stronger the evidence that all questions are measuring the same construct.

Coefficient alpha ordinarily ranges from zero to one. As a general rule of thumb, an alpha of .70 or higher indicates an acceptable level of reliability. However, the value of alpha is affected by factors such as test length. An examination with 300 questions will generate a higher alpha than a 100-item examination composed of similar questions. The level of alpha that indicates sufficient reliability also depends upon the kind of decisions to be made from test scores. For making high-stakes decisions about individuals (e.g., college admission), an alpha of .90—or even above .95—is desirable. It is often necessary to construct a very long test in order to attain a reliability coefficient this high. If a test is designed to inform decisions about groups (e.g., classrooms, instructors, schools), lower levels are acceptable. Lower values are also sufficient when the stakes are lower, for example, when the test score is just one component of determining one course grade. Coefficient alphas for classroom exams typically range from .50 to .90, rarely exceed .85, and are general considered acceptable if greater than .70.

The reliability coefficients (alphas) for the HAPS comprehensive examination, based on students taking the exam in 2015 are as follows: Red form, alpha=.90; Green form, alpha=.92.

These values are very high for a 100-item test. This level of reliability means the exam may be used both for self-evaluation of group performance (e.g., classrooms, instructors) and for making decisions regarding the knowledge levels of individual students.

The standard error of measurement (SEM) provides another way of looking at reliability. The SEM is a function of the reliability coefficient and the standard deviation of test scores. It can be used to create confidence intervals around a student's score. Human performance is not perfectly consistent; the same student is unlikely to get *exactly* the same score if tested again, due to various random factors. However, we would expect that 68% of the time, the student would score within 1 SEM of the current score, and 95% of the time, the same student would score within 2 SEMs of the current score. The HAPS SEMs are 4.4 for the Red form and 4.5 for the Green form. This gives us a 68% confidence interval of about 4 raw score points and a 95% confidence interval of about 9 score points for either form.

Conclusion: The HAPS A&P Comprehensive Exam exhibits solid evidence of validity as a measure of the HAPS learning outcomes. Validity is established not only via content and test development procedures but also in the exam's internal structure and consequential uses. Both forms of the HAPS A&P Comprehensive Exam also demonstrate high reliability and can be used to evaluate student performance as well as classroom performance and other grouped data.

Test Form Comparison

Table 2, below, shows summary statistics for scores on the Red and Green forms of the comprehensive examination. The two forms are quite comparable, although the Green form appears to be slightly easier than the Red form. Student performance was also slightly more variable on the Green form. Scores are reported to instructors as percentile ranks, allowing them to see their students' performance relative to the norm group for the form. Fewer students took the Green form in 2015, but the scores approach a normal distribution for both forms.

In comparison with classroom exams, the HAPS examination is a difficult test. However, it is not too difficult for high-performing students. The broad range of scores on both forms suggests that the exam is effective at assessing the full range of ability.

Table 2. Test Summary Statistics for Red and Green Forms

	Red	Green
N	1,840	217
Minimum	16	14
Maximum	97	100
Median	49	54
Mean	50.8	55.4
S.D.	14.1	16.3
SEM	4.4	4.5
Reliability	.90	.92

Neither form appears to have been speeded. Questions are designed to be concise, creating no more of a reading load than necessary. Time limits exceed the recommended minimum of 1 minute per multiple-choice question, and almost every student completes the entire exam; in fact, most students complete the exam well within the two hour maximum.

Conclusion: Student performance indicates that the two current forms of the HAPS A&P Comprehensive Exam are similar. Both measure a wide range of knowledge with a high degree of reliability.

Item Analysis

An examination is only as good as the individual questions, or items, it comprises. Questions on the HAPS comprehensive examination are reviewed by subject matter experts (SMEs), who evaluate the alignment of each question to the HAPS learning objectives as well as the degree to which each question conforms to psychometric item writing standards. In addition, the statistical performance of each question is evaluated. Standard psychometric practice includes evaluating test questions in terms of difficulty and discrimination. The most commonly used classical item statistics are the p-value for difficulty and the point-biserial correlation for discrimination.

The p-value is the proportion of test takers who answered a question correctly. It may be helpful to think in terms of easiness, rather than difficulty. The higher the p-value, the *easier* the question. For example, a question with a p-value of .90 was answered correctly by 90% of the students tested; it is a very easy question. On average, questions on both forms were somewhat difficult, yet there is a very wide range from extremely difficult (one item on the Red form was answered correctly by only 14% of test takers) to extremely easy (93% answered correctly).

Table 3. P-values (Item Difficulties) for the Red and Green Forms of the HAPS A&P Comprehensive Exam

	Red	Green
Mean	0.51	0.55
Minimum	0.14	0.24
Maximum	0.93	0.92

The two forms are similar in average item difficulty. Questions on the Green form may be slightly easier, but the p-value is affected by both the characteristics of the question and the group taking the test. The statistics shown here are based on a smaller number of students for the Green form. The modest difference in difficulty may disappear as more data become available for the Green form.

Average p-values of .50 to .55 are very reasonable, given the nature of the exam. These p-values suggest that the HAPS Comprehensive A&P Examination is more difficult than the typical classroom exam. This is to be expected because the HAPS exam is a national exam designed to align with the HAPS learning outcomes. Alignment with the local curriculum and classroom emphasis will vary, and some students in the norm groups are more motivated than others to succeed. Instructors will find test results most meaningful when their own students are motivated to do well—for example, when test scores are used as a factor in determining grades. At the same time, instructors should be aware of the national norms, consider the nature of the exam, and normalize the data in accordance with their own curriculum and student performance before incorporating scores into their grading algorithm.

The point-biserial correlation describes the relationship between scores (right or wrong) on each question and the total score on the examination. This serves as a measure of how well each question *discriminates* between students who have more knowledge and those who have less knowledge. As a correlation, the point-biserial can range from -1 to +1. In practice, however, it tends to range from small negative values (e.g., -.15) to moderate positive values (e.g., +.50). Higher values indicate better discrimination. A general rule of thumb is that a point-biserial of .20 or higher reflects very good discrimination, values of .15 or higher are reasonable, and lower values may be acceptable, depending on other characteristics of the question. Negative values are undesirable because they indicate that low-performing students tend to select the correct answer while the more knowledgeable students tend to select an incorrect answer.

The value of the point-biserial is somewhat dependent on the difficulty. Items that are very easy (or very hard) may have a low discrimination simply because there is very little variance in item scores. There is no room for the question to discriminate because nearly everyone is answering it correctly (or incorrectly). Thus, an examination with a range of difficulty is likely to include a few questions with low discrimination.

Both forms of the HAPS comprehensive examination show excellent discrimination. No questions have a negative discrimination. Very few have a point-biserial below .15, and the average is higher than is typically found in achievement testing. The overall pattern reflects an examination that discriminates very well between students with less knowledge and those with more knowledge of anatomy and physiology. Questions on the Green form are slightly more discriminating than those on the Red form, though this may be due to differences in the variability of the norm groups. The two forms may be regarded as generally comparable, especially when student performance is expressed as percentile ranks.

Table 4. Point-Biserials (Item Discriminations) for the Red and Green Forms of the HAPS A&P Comprehensive Exam

	Red	Green
Mean	0.27	0.32
Minimum	0.08	0.11
Maximum	0.46	0.55

Conclusion: HAPS A&P Comprehensive Exam questions demonstrate excellent discrimination, and the level of difficulty is appropriate for this type of exam.

Security

The HAPS comprehensive examination is administered online under secure conditions. Instructors are not permitted to review the examination and are required to assent to security guidelines. Examinations are password-protected. Passwords are provided to the proctors and are not available to students. Access to the exam is limited to specific individuals at pre-scheduled times. Each student is assigned a unique identifier. Students are not permitted to review questions they have already answered, making it more difficult for item harvesting to occur. Questions are randomized on each exam. In addition, the testing software is designed to track data that can detect unusual activity.

Industry-standard informational technology tools are in place to prevent hacking and block viruses. Student identity is protected by storing only their unique identifiers (i.e., no names). Such measures help to ensure the privacy of student data as well as protect the content of the examination.

Interpretation of Scores

Instructors should be aware that the HAPS A&P Comprehensive Exam is designed to test a broad body of knowledge that is not necessarily aligned with course curriculum. Student performance expressed as a raw score (i.e., number or percent correct) reflects mastery of the HAPS learning outcomes. The examination is, by design, more difficult than most classroom tests. Instructors should take this into account when using the HAPS exam as a final exam or otherwise including scores in grading and should adjust their grading scale accordingly. HAPS also reports percentile ranks, which most instructors will find informative in comparing the performance of their students to those of students in other classrooms around the country. Because there is a slight difference in difficulty between the Red and Green forms, comparisons will be most meaningful among groups taking the same form. Percentile ranks (not raw scores) should be used for any comparisons across the two forms. Instructors should note the description of the appropriate norm group in evaluating the extent to which their own students' performance meets expectations.

Uses for the HAPS A&P Comprehensive Exam

The HAPS Comprehensive A&P Examination has been used for a variety of purposes, including these:

- as a pre-test (on the first day/week of class) and again as a post-test (at end of course), to demonstrate the anatomy and physiology knowledge gained;
- as a placement exam; in other words, a demonstration of adequate student competency in anatomy and physiology (to substitute for retaking anatomy and physiology coursework);
- as a final exam to compare learning among students in the same class;
- as a self-evaluation tool for instructors desiring feedback on student learning relative to the HAPS learning outcomes;
- to assist instructors in evaluating effectiveness of teaching techniques;
- to measure learning outcomes in a newly developed course and compare with outcomes from traditional coursework;

- to provide normative data about the relative performance of one class of students to other students taking the exam during the same term;
- to obtain comparative data for benchmarking a class' student performance data relative to that of other A&P students across the United States;
- to evaluate and compare student performance in different A&P class sections, taught by different instructors (as a measure of learning consistency among different sections);
- to compare class performance, including pre/post-test gains, for the same course content presented in different modalities (e.g., online versus face-to-face instruction).

Future of the HAPS Comprehensive A&P Exam

In order to ensure the security of exam content and thorough mapping of exam questions to HAPS A&P Learning Outcomes, the HAPS testing task force is meeting (and will continue to meet) on a regular basis to create, edit, and vet new exam questions. A subset of newly written questions will be added to the current red and green versions of the HAPS A&P comprehensive exams. These newly written questions will not count toward a current exam grade, but rather, only psychometric data for the question will be collected. These data will be assessed by a professional psychometrician who will provide guidance to the task force about whether a question is considered reliable as is, or if the question needs to be edited. The task force will edit those selected questions and then the edited questions will be added back to the exam (again, not counting towards a student's exam score) to collect new psychometric data. Once a question's psychometric characteristics are acceptable, and the testing task force agrees that the question is a valid assessment of a student's A&P knowledge, the new question will be added to a growing database of questions to be used on future iterations of the HAPS A&P exam.

References

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

National Board of Medical Examiners (2002). *Constructing written test questions for the basic and clinical sciences* (3rd ed.). Philadelphia, PA: National Board of Medical Examiners. Available from *http://www.nbme.org/publications/item-writing-manual-download.html*.

Appendix A: HAPS Testing Task Force Members

Jennifer Burgoon, Ph.D. Assistant Professor, Division of Anatomy Department of Biomedical Education & Anatomy College of Medicine The Ohio State University	**Valerie O'Loughlin, Ph.D.** Professor of Anatomy and Cell Biology Medical Sciences Program Indiana University School of Medicine President Emeritus, Human Anatomy and Physiology Society *McGraw-Hill Textbook author of:* *McKinley/O'Loughlin/Pennefather-O'Brien: Human Anatomy 5e* *McKinley/O'Loughlin/Bidle: Anatomy & Physiology – An Integrated Approach 2e*
Curtis DeFriez, M.D., M.S. Professor Foundational Sciences Central Michigan University College of Medicine *Elsevier textbook chapter author of:* *Huether & McCance: Understanding Pathophysiology 5e*	**Kyla Ross, Ph.D.** Director of Graduate Training Department of Biomedical Engineering, Georgia Tech College of Engineering & Emory School of Medicine *McGraw-Hill Textbook author of:* *Eckel/Bidle/Ross: Human Anatomy and Physiology Lab Manual*
Kerry Hull, Ph.D. Professor , Department of Biology Bishops University (Canada) *Wolters Kluwer textbook author of:* *Cohen & Hull; Human Body Health and Disease* *Cohen & Hull: Structure & Function of the Human Body* *McConnell & Hull: Human Form Human Function*	**Dee Silverthorn, Ph.D., FAPS** Professor of Physiology Dell Medical School The University of Texas – Austin President Emeritus, Human Anatomy and Physiology Society *Pearson textbook author of:* *Silverthorn: Human Physiology: An Integrated Approach 7e*
Tom Lehman, M.S. Anatomy and Physiology Instructor Coconino Community College, Flagstaff, AZ President Emeritus, Human Anatomy and Physiology Society	**Eric Sun, Ph.D.** Associate Dean and Professor of Biology Middle Georgia State University
	John Waters, Ph.D. Lecturer, Department of Biology Penn State University President Emeritus, Human Anatomy and Physiology Society