American Psychological Association

Scholarship of Teaching and Learning in Psychology

ISSN: 2332-2101

2021, Vol. 7, No. 2, 123–139
http://dx.doi.org/10.1037/stl0000172

Development and Validation of an Introductory Psychology Knowledge Inventory

Erin D. Solomon, Julie M. Bugg, Shaina F. Rowell, Mark A. McDaniel, and Regina F. Frey
Washington University in St. Louis

Paul S. Mattson
Peninsula College

The introductory psychology course is offered by nearly all undergraduate psychology programs. Given its prominence in curricula, there have been calls to develop reliable and valid assessments of introductory psychology knowledge to inform and evaluate instruction. The current research, guided by the American Psychological Association's five content pillars model for the introductory course, developed and evaluated a relatively brief inventory that can be easily administered and scored to assess students' knowledge of introductory psychology. We followed established test development guidelines to write and revise inventory items. Then, we conducted two evaluation studies in which we administered the inventory to introductory psychology courses at two higher education institutions and to a sample of US adults to assess the reliability and validity of the test. Results suggested that the inventory has convergent (high correlations with course exam scores, all $ps < .001$), discriminant (low correlations with ACT English and Reading scores, all $ps < .05$), and known-groups validity (students that had taken high school psychology scored higher, all $ps < .001$). The inventory also demonstrated adequate test–retest reliability across a 1-week time interval, $r = .80$, $p < .001$. In a third study, we surveyed introductory psychology instructors about the inventory. The majority (97.4%) reported that it was representative of typical course content, and 75.7% reported they would use the inventory. We conclude that this introductory psychology knowledge inventory offers a practical, useful, reliable, and valid means of assessing students' knowledge and learning in the introductory course.

Keywords: assessment, introductory psychology, inventory, knowledge inventory

Supplemental materials: http://dx.doi.org/10.1037/stl0000172.supp

Nearly all undergraduate psychology programs offer an introductory course (Norcross et al., 2016), and across the U.S., introductory psychology courses collectively enroll between 1.2 and 1.6 million college students each year (Gurung et al., 2016). The already popular course will likely continue to grow in popularity given that medical and health disciplines are

This article was published Online First September 30, 2019.
Erin D. Solomon, Center for Integrative Research on Cognition, Learning, and Education (CIRCLE), Washington University in St. Louis; Julie M. Bugg, Department of Psychological and Brain Sciences, Washington University in St. Louis; Shaina F. Rowell and Mark A. McDaniel, CIRCLE and Department of Psychological and Brain Sciences, Washington University in St. Louis; Regina F. Frey, CIRCLE and Department of Chemistry, Washington University in St. Louis; Paul S. Mattson, Department of Psychology, Peninsula College.

Erin D. Solomon is now at the Bioethics Research Center, Washington University School of Medicine.

The authors would like to acknowledge and thank Heather Rice, Jan Duchek, Brian Carpenter, Emily Cohen-Shikora, and Heike Winterheld for feedback on the inventory and allowing data to be collected from their courses.

Correspondence concerning this article should be addressed to Julie M. Bugg, Department of Psychological and Brain Sciences, Washington University in St. Louis, St. Louis, MO 63130. E-mail: jbugg@wustl.edu

beginning to acknowledge the importance of psychosocial factors on behavior, as evidenced by the addition of the Psychological, Social, and Biological Foundations of Behavior section to the Medical College Admission Test (Mitchell, Satterfield, Lewis, & Hong, 2016).

The American Psychological Association (APA) has noted that the introductory course has the ability to shape the discipline (APA, 2014), and understanding the knowledge base in psychology is the first goal of the APA's guidelines for the undergraduate psychology major (APA, 2013). However, there is little research on what is typically taught in the course (APA, 2014). There have been calls for the creation of a common core of content and for the development of reliable and valid assessments of learning (APA, 2014; Dunn et al., 2010; Homa et al., 2013). These calls have led the APA to convene working groups to address these issues (APA, 2018). Development of a common core and reliable and valid assessments would provide some standardization across the course, and afford students and the general public a more unified message about what psychology is and how it can enhance understanding of human behavior (APA, 2014). Additionally, growing interest in university, departmental, and program assessment (e.g., Dunn, McCarthy, Baker, Halonen, & Hill, 2007) compels the need for reliable and valid measures of student learning.

The current study sought to contribute to these discussions by developing and validating an assessment for use in a variety of undergraduate introductory psychology courses. We first overview attempts to characterize the content of the introductory course and to assess learning gains of course content across an academic term. We then describe the development of an introductory psychology knowledge inventory and report three studies conducted to establish reliability, validity, and usefulness of the inventory. Studies 1 and 2 were focused on establishing that the inventory was sensitive to learning gains, had convergent and discriminant validity, and was reliable. Study 3 was focused on determining the usefulness of the inventory among a sample of experienced introductory psychology instructors.

There have been multiple attempts to assess the breadth and diversity of content in the introductory course. Several studies have gauged these characteristics by examining textbook content, and findings generally reveal a wide variety of terms in textbook glossaries (see Griggs, Bujak-Johnson, & Proctor, 2004; Nairn, Ellard, Scialfa, & Miller, 2003; Proctor & Williams, 2006; Zechmeister & Zechmeister, 2000). For example, Proctor and Williams (2006) found that only 33 terms out of 4,902 total terms were common to all 33 textbooks they analyzed, while 428 terms were found in at least half of the textbooks. Gurung et al. (2016) reviewed this line of research and concluded that there is broad consistency in the 14–16 general topics covered in the introductory course, but far less consistency in the specific content covered within these general topics. Taking a different approach, Homa et al. (2013) examined 158 introductory psychology syllabi to determine what content areas were taught, and how much course time was spent on each content area. They found that instructors spent the most time teaching cognitive and physiological content (nearly 20% of course time per topic); moderate amounts of time on social, clinical, developmental, and "other" content (roughly 8%–13% of course time per topic); and the least time on research methodology and history (less than 5% of course time per topic; Homa et al., 2013).

The APA convened a working group to address the common core issue, and as a result outlined five content pillars that are core to the introductory psychology course: biological, cognitive, development, social and personality, and mental and physical health (APA, 2014). They recommended that every introductory course covers the five pillars, and suggested instructors move away from a "silo" model in which content areas are treated as separate subfields (APA, 2014). Most importantly for present purposes, in addition to the calls for more content consistency in the introductory course, the APA and others have called for the development of knowledge assessments for the introductory course (APA, 2014; Homa et al., 2013). Having reliable and valid assessments of knowledge is important for multiple reasons, including alerting students and instructors to difficult topics prior to instruction and helping instructors better understand the effectiveness of their teaching practices (Adams & Wieman, 2011; Bass, Drits-Esser, & Stark, 2016; Libarkin, 2008). In many cases, it is not sufficient to use course exams for this purpose, because

course exams can be idiosyncratic, may change from year to year, and their reliability and validity are generally unknown. Furthermore, in the context of educational research, using exams as the measure of learning makes it difficult to compare research results from multiple studies.

Knowledge Inventories

A common approach for assessing learning in educational contexts is through the use of knowledge inventories. These inventories aim to assess learning gains made over the course of the term by using a pre/post design. Many science fields have developed inventories for their introductory courses, which are used in formal studies to evaluate the effectiveness of various teaching strategies and pedagogical innovations (e.g., Epstein, 2013; Hestenes, Wells, & Swackhamer, 1992; Garvin-Doxas & Klymkowsky, 2008; Smith, Wood, & Knight, 2008). For example, the development of the Force Concept Inventory in physics led to a wealth of physics education research studies in the years following, including research using this inventory to establish that "interactive engagement" in classrooms is generally more effective for learning general physics than is traditional lecture (Cahill et al., 2014; Hake, 1998). The field of psychology in particular has been identified as one that could benefit from the development of an inventory for the introductory course (Hake, 2015). Along these lines, a team of psychologists recently developed and validated an inventory for psychological research methods courses, called the Psychological Research Inventory of Concepts (Veilleux & Chapman, 2017a, 2017b). This initial effort is valuable, because research methods is also a course that is common to most psychology curricula (Norcross et al., 2016).

There are some existing assessments measuring general psychology knowledge. The Education Testing Service (ETS) provides both the Major Field Test for Psychology and the GRE Subject Test in Psychology (ETS, 2018a, 2018b). Both assessments are roughly 200+ questions and need to be purchased from ETS. Unfortunately, the length and the financial burden associated with these tests do not make them an ideal or an attractive option for most instructors. Another assessment that requires purchasing is the Psychology Area Concentration Achievement Test (i.e., ACAT-P; PACAT, Inc., 2018). The ACAT-P covers up to 12 psychology content areas, and instructors select which content areas they would like to administer to their students. The number and format of items are unpublished, but time estimates range from 48 min to complete four content areas to 120 min to complete 10 content areas (PACAT, Inc., 2018). An important limitation of these assessments for present purposes, however, is that they are generally administered at the end of the undergraduate curriculum to gauge achievement (although the ACAT-P is often administered additionally during students' first year in college to assess pre/post change). Given that these are aimed for senior level students, test difficulty and breadth could be a deterrent for administering them as a pre/post assessment of knowledge gained in an introductory course.

Additionally, there are a few assessments that are perhaps better suited for assessing introductory psychology knowledge than the assessments described previously. A test developed by Peter, Leichner, Mayer, and Krampen (2015) aimed to be a short measure of basic knowledge in psychology and consisted of 21 items covering the concepts that are most commonly found in introductory psychology textbooks. The assessment was developed and validated in German, but an English translation is available. It has multiple types of questions, including multiple-true/false, fill in the blank, matching, and open ended. While not specifically designed for the introductory course, the fact that the content of the test was drawn from introductory texts perhaps makes it appropriate for use in introductory courses.

In another effort, Thompson and Zamboanga (2004) developed measures of psychology knowledge and popular psychology myths in their work examining prior knowledge and academic achievement in the introductory psychology course. Their psychology knowledge measure consisted of 25 multiple-choice items that were based on faculty-identified central concepts, issues, or ideas introductory students should know. Their popular psychology myths measure consisted of 16 statements rated on a Likert-type scale ranging from *very sure it's false* to *very sure it's true*. The statements centered on psychological ideas about which lay

individuals tend to have strong intuitions, even though psychological research does not support the ideas.

All three of these assessments (i.e., the measure from Peter et al., 2015, and the two measures from Thompson & Zamboanga, 2004) are short in length and more closely aligned with the introductory course than those previously described. However, given the broad nature of introductory psychology curricula, there is likely not a one-size-fits-all assessment that would work for the majority of introductory psychology courses. For example, Peter et al.'s (2015) scoring is not intuitive due to the multiple types of response options. Additionally, Thompson and Zamboanga's (2004) measures were, as far as we know, not developed using traditional test-development procedures. Thus, there is a need for another assessment that is both reliable and valid.

Development of an Introductory Psychology Knowledge Inventory

We sought to develop a knowledge inventory that was reliable and valid, and additionally was relatively short, easy to administer, easy to score, and relatively difficult so that learning gains would not be obscured by ceiling effects. We centered the content of the inventory on past work defining the common core of introductory psychology knowledge, including the APA's recommendations (e.g., APA, 2014).

Item Development

We followed established item-development guidelines to develop the inventory items (e.g., Haynes, Richard, & Kubany, 1995). Generally, our process of item development was iterative, with feedback from various experts along the way.

First, we sought to identify the important facets[1] of introductory psychology knowledge. We gathered information from past literature by examining multiple introductory psychology texts and discussing facets among authors and colleagues. Through this analysis and discussion, we identified 14 content areas that embodied the important facets of introductory psychology knowledge. The 14 content areas were research methods, biological foundations, sensation and perception, consciousness, learning, memory, thinking, intelligence, life span development, emotion, health and well-being, personality, psychological disorders and therapy, and social psychology. These content areas fit within the APA's five pillars structure (see Table 1). Although these content areas are broad in scope, it is worth noting that they do not represent the entire field of psychology; for example, the APA has more than 50 divisions. Nonetheless, we thought these 14 content areas represented the content of many introductory courses, as evidenced by past literature and discussions with introductory psychology instructors.

From these 14 content areas, one of the authors (J.B., who was one of the instructors teaching the course in Sample 2 in Study 1) wrote between one and four multiple choice items per content area, basing the number of items on the breadth and importance of the content area in a typical introductory psychology course (e.g., Homa et al., 2013). For example, 10 items were written that broadly fall under the Cognitive APA pillar, while only three items were written for the Development pillar.

Next, multiple experts reviewed the items as a way of assessing content validity. First, psychology faculty and research staff from one of the institutions in Study 1 (i.e., the institution for Samples 1 and 2) reviewed the items. All ($n = 6$) were current or former instructors of introductory psychology courses. Feedback was provided to the research team, and problematic items were revised or replaced.

Second, we sought feedback from psychology faculty outside the institution ($n = 4$ faculty representing four different institutions). We administered an online survey assessing a) whether the 14 content areas were representative of the content covered in introductory psychology (i.e., were all important facets accounted for), b) whether the differing numbers of items for each content area were representative of the importance of that content area to introductory psychology, and c) whether any of the items were redundant. All respondents indicated the 14 content areas were representative

[1] The term "facet" is used in scale and test development literature (e.g., Haynes et al., 1995) and refers to the specific parts or components of the construct being measured.

Table 1
Categorization of the Inventory's 14 Content Areas Within the APA's Five Pillars of the Introductory Psychology Course

APA Pillar	Introductory Psychology Knowledge Inventory Content Area
Biological	biological foundations, sensation and perception, consciousness
Cognitive	learning, memory, thinking, intelligence
Development	lifespan development
Social and Personality	personality, social psychology, emotion
Mental and Physical Health	health and well-being, psychological disorders and therapy

Note. The content area of "research methods" does not fit within any one of the five pillars, because the model indicates that research methods should be included in each of the five pillars.

of typical introductory psychology courses and that no facet was missing, that the number of items written for each content area was appropriate, and that no items were redundant. Additionally, we asked whether each item was clearly worded, which content area(s) they thought it assessed, and whether they had any specific feedback on that item. Based on feedback from this survey, we continued revising items, and replaced three items.

Last, we piloted the items with undergraduate students ($n = 3$): two psychology majors and one economics major. We sought student feedback on difficulty, clarity, and how long the overall inventory took to complete. All three students indicated that the inventory was difficult and provided feedback on item clarity. Additionally, they reported that the inventory took approximately 22–25 min to complete.

As a result of our development activities, the initial inventory consisted of 32 multiple-choice items, each with five response options. Our next steps involved conducting three studies to assess the difficulty of the inventory, to assess the validity and reliability of the inventory, and to gather additional feedback from psychology faculty. We sought to collect multiple samples from different course and institution types to investigate whether the difficulty of the test was appropriate in different contexts. We also sought to determine whether the inventory had convergent, discriminant, and known-groups validity, as well as internal consistency and test–retest reliability. Additional item-level analyses were conducted and can be found in the supplemental materials.

Study 1

In Study 1, we administered the inventory to six introductory psychology courses. We assessed the difficulty of the inventory in all six of the courses, and for two courses (Samples 1–2), we collected additional data that allowed us to assess whether the inventory could detect learning gains. Last, we investigated whether there was evidence of convergent (Samples 1–2), discriminant (Samples 1–2), and known-groups validity (all six samples), and internal consistency (all six samples).

First, we sought to assess whether the inventory had appropriate difficulty and could detect learning gains. Designing a test that is difficult at pretest provides information to instructors regarding what type of knowledge their students start the class with, so that instruction can accommodate prior knowledge (or lack thereof). Additionally, if the test is administered at both the beginning and the end of the term, designing a test that is difficult reduces the risk of ceiling effects at posttest, which would constrain the data and potentially miss important learning gains. Thus, we aimed for the difficulty of the assessment to be between 20% and 40% correct at pretest, and roughly 60% correct at posttest, on average. Additionally, we calculated learning gains by examining the normalized gain score, a common practice for other STEM knowledge inventories (e.g., Hake, 1998). We aimed to assess the inventory's difficulty across multiple types of institutions and introductory courses, to assess its utility in different settings. In particular, we aimed to gather data from both in-person and online introductory courses, given the move toward online administration of courses in recent years.

Second, we aimed to assess the convergent, discriminant, and known-groups validity of the inventory. Broadly, these three types of validity fall under the umbrella of construct validity, which is the idea of whether a test measures what it claims to measure. Convergent validity is the degree to which the test is related to another measure that purports to measure the

same or a similar construct (Campbell & Fiske, 1959). Evidence that a test has convergent validity would be a significant positive correlation with the other measure. To gauge convergent validity of the inventory, in the present study we examined the correlations between posttest scores and exam performance in two of the introductory courses we sampled (Samples 1–2). Theoretically, both of these measures are assessing introductory psychology knowledge; therefore, if they are correlated, this would demonstrate convergent validity.

Similar to convergent validity, discriminant validity also deals with the relationship between the test and other measures. In contrast to convergent validity, however, discriminant validity is the idea that the test would not be related to another measure with which it should be theoretically unrelated (Campbell & Fiske, 1959). Evidence of discriminant validity would be small or nonsignificant correlations with the other measure. In this study, we examined whether the inventory was correlated with measures of English and reading proficiency (Samples 1–2). A low correlation between the inventory and English/reading proficiency would show that the inventory has discriminant validity, at least in terms of discriminating psychology knowledge from general verbal ability.

Known-groups validity is when two or more groups that you would expect to differ on the construct being assessed, are shown to actually differ on the construct (Cronbach & Meehl, 1955; Hattie & Cooksey, 1984). Demonstrating that the groups differ significantly on the construct would be an indicator that the measure is measuring the intended construct. In the current study, we examined whether inventory scores at the start of the courses (i.e., on the pretest) would differ for students that had previously taken a psychology course (all six samples). Known-groups validity would be established if a group of students with more prior psychology knowledge at the outset of the course (e.g., students who had taken a high-school psychology course) displayed significantly higher inventory scores than a group of students with less prior psychology knowledge (e.g., students who had not previously taken a psychology course).

Finally, we examined reliability via internal consistency reliability, which is the degree to which all items on a test measure the same construct (Henson, 2001). It is typically measured by calculating Cronbach's alpha coefficient, which is essentially the average interitem correlation (Cortina, 1993; Cronbach, 1951; Henson, 2001). While internal consistency may not be wholly appropriate for evaluating the reliability of the present inventory given that one may not expect a student to perform similarly in all of the 14 content areas, reporting of Cronbach's alpha is common practice in measurement development (Henson, 2001). Generally, an alpha of above .70 would indicate adequate internal consistency reliability (Cortina, 1993).

Method

Participants. We administered the inventory to six introductory psychology courses at two institutions. Students signed informed-consent forms authorizing the release of study-related data to the research team.

Samples 1 and 2. The first and second samples were drawn from two high-enrollment (~250–500 total students each) introductory psychology courses taught in the fall 2016 and spring 2017 terms, respectively, at a selective research university in the Midwestern United States. In both courses, each of three instructors taught two sections of the course for roughly five consecutive weeks (for a total of 15 weeks for the term), which was a typical model for this course at this institution. The three instructors differed between the two terms (i.e., there were six instructors in total). Students in both samples earned a small amount of extra credit in the course for participating in the study. The courses were taught in a single term and were not part of a sequence (e.g., taught across semesters or quarters), and covered all 14 content areas of the inventory.

Samples 3–6. The third, fourth, fifth, and sixth samples were drawn from four small (~25–60 students) introductory psychology courses at an open-access community college in the Northwestern United States. Each of the courses was 10 weeks long and taught by the same instructor. The courses in Samples 3 and 4 were taught during the winter 2017 term, and the courses in Samples 5 and 6 were taught in the spring 2017 term. The courses in Samples 3 and 5 were taught in person, while the courses in Samples 4 and 6 were taught online. Students completed the inventory as part of the course,

and only data from consenting students were released to the research team. The inventory was administered only at the start of the course in these samples (i.e., pretest). Like Samples 1–2, the courses were taught in a single term and were not part of a sequence. However, 10 of the 14 content areas of the inventory were covered in the courses.

Measures.

Knowledge inventory. The knowledge inventory was administered at both the start and end of the term (i.e., pretest and posttest) for Samples 1 and 2, and at the start of the term for Samples 3–6 (i.e., pretest only). The inventory items differed slightly between samples due to item refinement. Specifically, in Sample 2, three new items replaced three poorly performing items (i.e., Sample 1 and Samples 3–6 had the same items). Inventory items were administered outside of class time using an online survey platform. Additionally, in Samples 1 and 2, items were administered in a random order that differed for every student. We scored the inventory by calculating a proportion of items answered correctly.[2]

Exam performance. We collected exam performance data for students in Samples 1 and 2. Even though different instructors taught the two courses and all the exam questions differed, the general exam structure was the same between the two samples. Specifically, there were three noncumulative unit exams, which were administered at the end of each 5-week instructional unit. All three of these unit exams consisted of 50 multiple-choice questions and were written by the instructor who taught the content for that unit. Additionally, there was a cumulative final exam administered at the end of the term, which also consisted of 50 multiple-choice questions. Each instructor wrote approximately ⅓ of the questions for the final exam. Students were informed at the start of the term that their lowest exam score (out of the four exams: three unit exams and one final exam) would be dropped when calculating their final grade. About ⅓ of students in each course opted not to take the final exam. Our measure of exam performance was therefore the average of the three highest (of the four total) exam scores.

ACT scores. For Samples 1 and 2, we retrieved students' ACT English and Reading scores from the university. The majority of students at the institution completed the ACT test, but for those students with only SAT scores we converted their SAT scores to an ACT equivalent score (see Dorans, 1999). Because of this conversion, we utilized ACT English and Reading scores (i.e., ACT ER) together as one score, which was equivalent to the SAT Verbal score. Scores on this ACT ER measure could range from 0 to 36. Neither ACT nor SAT information was available for 2 students in Sample 1 and 10 students in Sample 2.

High-school psychology. After completing the inventory, students in all six samples self-reported whether they had taken a psychology course in high school. In Samples 1 and 2, 95/316 (30%) and 42/148 (28%) students completed a psychology course in high school, respectively. Because there were so few students in Samples 3–6 that had completed a high school psychology course, we combined those four samples for analyses using these data. Once combined, only 17/114 (15%) of students in Samples 3–6 completed a psychology course in high school.

Demographics. For Samples 1 and 2, we retrieved students' gender, race, and academic level (i.e., first-year, sophomore, junior, senior) from the institution. For Samples 3–6, we retrieved students' gender, race, and age from the institution.

Results

The majority of students in each of the six courses consented to participate in the study: Sample 1: 363/467 students (78%), Sample 2: 208/274 (76%), Sample 3: 24/30 (80%), Sample 4: 38/58 (66%), Sample 5: 28/34 (71%), and Sample 6: 23/44 (52%). We removed students who were completing the course for pass/fail credit, because pass/fail students are known to have different course motivations than other students (Brownell et al., 2015; Sample 1: $n = 36$, Sample 2: $n = 57$, and Samples 3–6: $n = 0$).[3]

[2] With Samples 3–6 combined, the median amount of time to complete the inventory was 22.00 minutes. We did not have equivalent timing information for Samples 1 and 2. However, inventory data collected from later samples at that institution showed that the median amount of time to complete the inventory was 15.17 minutes.

[3] When the analyses were conducted including the pass/fail students, all study results except one were the same. The one finding that changed was in Sample 2: The inventory and ACT ER were correlated at pre, $r(196) = .19$, $p = .007$.

Participants with missing data on more than one item on the inventory were excluded from all analyses (Sample 1: $n = 11$, Sample 2: $n = 3$, and Samples 3–6: $n = 0$). However, in an effort to retain participants, participants with missing data for only one item on the inventory were retained, and their missing response was marked as incorrect (Sample 1: $n = 7$, Sample 2: $n = 5$, Sample 3: $n = 0$, Sample 4: $n = 2$, Sample 5: $n = 1$, and Sample 6: $n = 0$).[4] Lastly, there were two students appearing in both Samples 1 and 2, indicating they were retaking the course for a second time; thus, we removed the two students from Sample 2. Final samples sizes are listed in Table 2.

Inventory difficulty and learning gains. We assessed the difficulty of the inventory by examining the proportion correct. As seen in Table 2, mean difficulty ranged from .29 to .41 at pre, and .62 to .64 at post. Importantly, all of the means at pretest were significantly different from chance (i.e., .20; Sample 1: $t(315) = 29.92$, $p < .001$; Sample 2: $t(147) = 19.48$, $p < .001$; Sample 3: $t(24) = 6.07$, $p < .001$; Sample 4: $t(37) = 5.51$, $p < .001$; Sample 5: $t(27) = 5.33$, $p < .001$; and Sample 6: $t(22) = 5.59$, $p < .001$). Given these results, it seems that the inventory had sufficient difficulty, in that scores were above chance at pre, but there was still substantial room for growth at post (i.e., no ceiling effects in Samples 1 and 2, which were the only samples with posttest scores).

Additionally, for Samples 1 and 2 we examined learning gains made over the course of the term. To examine this, we calculated the normalized gain, which represents the average amount students learned divided by the amount they could have learned, and is calculated using the formula (post−pre)/(1−pre) (Hake, 1998). The normalized gain was $M = .39$ ($SD = .22$) and $M = .37$ ($SD = .24$) for Samples 1 and 2, respectively. Additionally, for both Samples 1 and 2, we found that the change from pretest to posttest was significant, Sample 1: $t(315) = -29.12$, $p < .001$, Sample 2: $t(147) = -18.00$, $p < .001$. We can reasonably assume that students' knowledge of psychology would increase after completing the course; thus, this is evidence that the inventory may be successfully assessing learning gains.

Convergent validity. To assess convergent validity, we examined the correlations between post-inventory scores and exam averages in Samples 1 and 2 (the only samples for which the inventory was administered at posttest). Correlations were $r(314) = .44$, $p < .001$ for Sample 1, and $r(146) = .46$, $p < .001$ for Sample 2. These significant correlations are evidence of convergent validity.

Discriminant validity. To assess discriminant validity, we examined correlations of the inventory with ACT ER scores. If the inventory has discriminant validity, we would expect that inventory scores would have small or nonsignificant correlations with ACT ER. As a reminder, ACT data were only available for Samples 1 and 2. For Sample 1, we found that the inventory and ACT ER were significantly correlated both at pre, $r(312) = .13$, $p = .020$, and at post, $r(312) = .25$, $p < .001$. For Sample 2, the inventory and ACT ER were not correlated at pre, $r(136) = .14$, $p = .096$, but were correlated at post, $r(136) = .21$, $p = .016$. While three of these four correlations are significant, the magnitude of the correlations is small, which suggests there is only a weak relationship between the inventory and ACT ER, which is indicative of discriminant validity.

Known-groups validity. We examined known-groups validity by determining whether students who had completed a psychology course in high school performed better than students who had not completed a psychology course in high school. As a reminder, because there were so few students in Samples 3–6 who had completed a high-school psychology course, we combined those four samples for this analysis.

For two of the three t tests conducted on pretest scores, we found that students who had taken psychology in high school had higher scores than students who had not completed a psychology course in high school, Sample 1: $t(314) = -7.68$, $p < .001$ (high school $n = 95$, $M = .48$, $SD = .13$; no high school $n = 221$, $M = .37$, $SD = .11$); Sample 2: $t(146) = -3.95$, $p < .001$ (high school $n = 42$, $M = .47$, $SD = .12$; no high school $n = 106$, $M = .38$, $SD =$

[4] When the analyses were conducted using listwise deletion as the method of handling missing data, all study results except one were the same. The one finding that changed was that at posttest, there was a significant difference between students who had ($M = .67$, $SD = .14$) versus had not ($M = .63$, $SD = .14$) taken psychology in high school, $t(304) = -2.08$, $p = .038$.

Table 2
Characteristics of each Sample

Variable	Sample 1	Sample 2*	Sample 3	Sample 4	Sample 5	Sample 6
Term	Fall 2016	Spring 2017	Winter 2017	Winter 2017	Spring 2017	Spring 2017
Institution type	R1	R1	CC	CC	CC	CC
Course type	In-person	In-person	In-person	Online	In-person	Online
N	316	148	24	38	28	23
Pre-inventory (*SD*)	.41 (.13)	.41 (.13)	.36 (.13)	.32 (.13)	.29 (.09)	.31 (.09)
Post-inventory (*SD*)	.64 (.14)	.62 (.17)	—	—	—	—
Normalized gain of inventory (*SD*)	.39 (.22)	.37 (.24)	—	—	—	—
Exam average (*SD*)	.84 (.08)	.84 (.09)	—	—	—	—
ACT English + Reading (*SD*)	33.37 (2.11)	33.59 (2.03)	—	—	—	—
% first years	68.67	66.89	—	—	—	—
% Female	63.29	62.16	40.00	73.68	82.14	73.91
% White	54.09	43.92	68.00	60.53	60.71	56.52
Age (*SD*)	—	—	26.08 (12.86)	25.03 (8.26)	26.21 (8.93)	20.48 (4.65)

Note. R1 = selective research-intensive institution; CC = open-access community college; "—" = not applicable or not available; % first years = percentage of first-year students. Exam average, pre-inventory, and post-inventory scores are listed as proportion of items answered correctly. Numbers in parentheses are standard deviations. All numbers in the table are means unless otherwise indicated.
* Sample 2 inventory items differed slightly from the other samples, such that three items in Sample 2 replaced problematic items used in the other samples.

.12); Samples 3–6: $t(112) = -.721$, $p = .472$ (high school $n = 17$, $M = .34$, $SD = .16$; no high school $n = 97$, $M = .32$, $SD = .11$). These pretest findings from Samples 1–2 provide evidence of known-groups validity, whereas the finding from Samples 3–6 does not. Interestingly, for posttest scores, there were no differences in performance between students who had versus had not taken psychology in high school: Sample 1: $t(314) = -1.33$, $p = .184$ (high school $M = .66$, $SD = .15$; no high school $M = .64$, $SD = .14$); and Sample 2: $t(146) = -.51$, $p = .608$ (high school $M = .64$, $SD = .16$; no high school $M = .62$, $SD = .17$). This indicates that students who did not complete a psychology course in high school were performing just as well as their peers by the end of the course.

Internal consistency reliability. We calculated Cronbach's alpha as an indicator of the inventory's internal consistency, both for the inventory overall and for each section in the inventory. Higher Cronbach's alpha scores indicate higher internal consistency. Generally, the inventory as a whole was below the .70 standard threshold at pretest (Sample 1 $\alpha = .61$, Sample 2 $\alpha = .62$, Samples 3–6 $\alpha = .56$), but exceeded the standard at posttest (Sample 1 $\alpha = .73$, Sample 2 $\alpha = .81$). Regarding separate sections of the inventory, all Cronbach's alpha scores for separate inventory sections were below the .70 standard threshold at both pretest and posttest.

Discussion

Results from Study 1 supported the notion that the inventory appears to be an appropriate assessment tool for measuring introductory psychology knowledge. Specifically, results suggested that the inventory was sufficiently difficult in multiple types of introductory courses, including courses from a selective research-intensive institution and an open access community college. The pretest scores were sufficiently low (but not at floor) and the posttest scores were not at ceiling. Furthermore, we found that pretest scores in the online courses were similar to their in-person counterparts (i.e., Samples 3–6 all had similar difficulty scores). For the two samples in which we were able to examine learning gains (i.e., Samples 1 and 2), we found that students made significant gains in both courses, as measured by the inventory.

Regarding validity, we found evidence of convergent, discriminant, and known-groups validity. When administered at the end of the course (in Samples 1–2 only), inventory scores were correlated with course exam scores. Given the conceptual similarity that the inventory and

course exams share, this is evidence of convergent validity. Additionally, regarding discriminant validity, we found that inventory scores correlated only slightly with ACT ER. This suggests that the inventory is likely not simply measuring general English or reading ability. Lastly, for two of our samples (Samples 1 and 2), students who had completed a prior psychology course in high school performed better on the inventory at pretest than students who had not completed a high school psychology course. This is evidence of known-groups validity. This difference was not observed for Samples 3–6 (combined), which may reflect low power. A post hoc power analysis using G*Power (Faul, Erdfelder, Lang, & Buchner, 2007) indicated that we had only a 47% chance of detecting a medium-sized effect, given the sample size. An additional consideration is that the mean age of students in Samples 3–6 was 24.53 years, compared to 18–20 years in Samples 1–2 (i.e., estimated based on typical age of traditional college first years at the institution). Given that long-term retention of academic material is generally poor (e.g., Butler & Roediger, 2007; Landrum & Gurung, 2013; Larsen, Butler, & Roediger, 2009), it is possible that any prior knowledge of psychology was lost between high school and college for students in Samples 3–6. Indeed, Samples 3–6 were older than traditional college students, and it is therefore likely that more time had elapsed between high school and college for these samples relative to Samples 1 and 2.

Lastly, we examined the internal consistency of the inventory. We found that when administered at pretest, Cronbach's alpha scores were below the standard of $\alpha = .70$, but above $\alpha = .70$ at posttest. As outlined previously, we may not expect high internal consistency from this type of measure. For example, students' knowledge of one content area, such as biological psychology, may differ considerably from their knowledge of another content area, such as social psychology (McNamara, Williamson, & Jorgensen, 2011; Peck, Ali, Levine, & Matchock, 2006). Additionally, Cronbach's alpha conducted on separate sections of the inventory were all below the standard of $\alpha = .70$. One likely reason for the low alphas is that most sections contain only 2–3 items, and the Cronbach's alpha statistic favors scales with more items (Cronbach, 1951). Another issue potentially contributing to low alphas is that the diversity of content within each section is high, which would make a low alpha expected (e.g., a student may have good understanding of classical conditioning but poor understanding of operant conditioning, resulting in a low alpha for the learning section). Because Cronbach's alpha may not be an appropriate test of reliability for the inventory, we conducted a second study in order to better assess the reliability of the inventory.

Study 2

The primary purpose of our second evaluation study was to assess the test–retest reliability of the inventory. Test–retest reliability involves administration of the test at two different time points to determine the stability of the test over time (Geisinger, 2013). Adequate test–retest reliability would be indicated by a correlation above .80 between the two administrations of the test (Nunnally, 1978). A second purpose of Study 2 was to conceptually replicate our known-groups validity finding from Study 1. However, instead of comparing inventory performance for those who had and had not completed a high-school psychology course (which, as previously discussed, might be problematic for individuals for whom high school was many years in the past), we examined inventory performance as a function of number of college-level psychology courses completed. We assessed test–retest reliability and known-groups validity by administering the inventory twice to a sample from the general public, with the second administration occurring one week after the first.

Method

Participants. Data were collected from Mechanical Turk (mTurk), which is an online service through Amazon.com in which individuals can complete surveys and tasks for small amounts of money. In the current study, participants were paid $1 USD to participate, and we limited our data collection to mTurk workers residing in the United States. The average age of the sample ($N = 95$) was $M = 34.23$ years ($SD = 9.64$), 56.8% of the sample identified as female, and 77.9% identified as White. Regarding highest level of education completed, 10.5%

of the sample had completed high school, 18.9% had some college, 11.6% had an associate's degree, 45.3% had a bachelor's degree, 8.4% had a graduate degree, and 5.3% had a professional degree.

Measures.

Knowledge inventory. The inventory was administered at two time points, spaced one week apart. Inventory items, which were the same as in Sample 2 in Study 1, were again administered using an online survey platform. Items were administered in a random order that differed for every participant, and we scored the inventory by calculating a proportion of items answered correctly.

College-level psychology courses. At the end of the first administration of the inventory, participants were asked, "How many psychology courses did you complete in college?" with response options of "0," "1," "2," "3+," and "not applicable: did not attend college." Twenty participants responded that they completed "0" college psychology courses, 39 indicated "1," 11 indicated "2," 11 indicated "3," and 14 indicated not applicable.

Demographics. At the end of the first administration of the inventory, we collected information on participants' age, gender identity, race, level of education, and whether English was one of their native languages. During the second administration of the inventory, we asked participants whether they had studied or read about anything relating to psychology in the past week. Response options consisted of "yes," "no," and "unsure" with an open response area to describe what they had read or studied in the past week (if anything).

Results

One hundred fifty-one participants completed the first administration of the inventory, and 103 completed the second administration. After removing the participants who did not complete the second administration ($n = 48$, or a 68% retention rate), we additionally removed participants who were non-native English speakers ($n = 3$) or who indicated they had read about psychology or studied during the one-week testing interval ($n = 2$). Additionally, similar to Study 1, participants with missing data on more than one item on the inventory were excluded from the analyses ($n = 3$).[5] Participants with missing data for only one item on the inventory were retained, and their missing response was marked as incorrect ($n = 5$). This resulted in a final sample of 95 participants.

Test–retest reliability. We assessed the test–retest reliability by calculating the correlation between the two administrations of the test. Results showed that the two administrations, spaced one week apart, were significantly correlated, $r(93) = .80$, $p < .001$. The high magnitude of the correlation suggests that the inventory has sufficient test–retest reliability.

Known-groups validity. We examined known-groups validity by determining whether participants who completed more college-level psychology courses performed better than participants who completed fewer college-level psychology courses. For these analyses, we excluded anyone indicating they did not attend college ($n = 14$), leaving 81 participants in the analyses.

A mixed model analysis of variance (ANOVA) with the number of college psychology courses completed ("0," "1," "2," vs. "3+" courses) as the between-subjects variable and time (first administration vs. second administration) as the within-subjects variable was conducted. There was a main effect of number of courses, indicating that participants who had completed more college-level psychology courses performed better on the inventory: $F(3, 77) = 6.11$ $p < .001$, generalized $\eta^2 = .173$. Tukey's HSD post hoc analyses showed that participants indicating they had completed 3+ college psychology courses ($M = .45$, $SD = .12$) performed higher on the inventory than participants completing zero courses ($M = .28$, $SD = .11$, $p < .001$), one course ($M = .32$, $SD = .13$, $p = .002$), and two courses ($M = .33$, $SD = .11$, $p = .009$). There was no main effect of time, indicating that there were no significant differences on inventory scores between the first and second administrations: $F(1, 77) = .61$, $p = .43$, generalized $\eta^2 < .001$ (i.e., no practice effect). Additionally, the interaction between number of college psychology courses and time was not significant, $F(3, 77) = .55$, $p = .65$, generalized $\eta^2 = .002$. These results support that the inventory has known-groups validity.

[5] When the analyses were conducted using listwise deletion as the method of handling missing data, all study results were meaningfully the same.

Discussion

Results from Study 2 indicate that the inventory has adequate test–retest reliability, and participants do not appear to be gaining much knowledge based on the experience of taking the inventory (i.e., no evidence for a practice effect). We also found converging evidence for known-groups validity by conceptually replicating the known-groups validity findings from Study 1. Specifically, we found that participants who had completed more psychology courses in college (i.e., 3+ courses) scored higher on the inventory than those who had completed fewer or no college psychology courses.

Study 3

The primary purpose of our final evaluation study was to gather additional feedback from a larger group of introductory psychology instructors than we had polled prior to Studies 1 and 2. Specifically, we sought qualitative and quantitative feedback on the inventory items, whether the items were representative of a typical introductory psychology course, and how likely instructors were to use the inventory.

Method

Participants. Data were collected from introductory psychology instructors ($N = 38$) recruited via snowball convenience sampling through the authors' professional networks, contacting attendees of a psychology teaching conference, and asking participants to forward the recruitment materials to their professional networks. Participants could enter a raffle to win one of two $50 gift cards for participating.

Participants were from at least 30 different institutions, with six participants not reporting their institution. Seventeen (44.7%) were from private four-year institutions, 13 (34.2%) were from public four-year institutions, 6 (15.8%) were from community colleges, and 2 (5.3%) were from high schools. Participants reported having taught the introductory psychology course an average of 20.89 times ($SD = 26.96$); when two outlier responses were removed (responses of 95 times and 150 times), the average was 15.09 times ($SD = 9.13$).

Measures. Participants completed an online survey. First, the survey contained all the knowledge inventory items, for which participants were asked an open-ended question: "Are there any items that are problematic? If so, please list the item number(s) that are problematic with a short explanation of the problem." This yielded a variety of qualitative data. Second, participants were asked a number of questions regarding the inventory and whether they would use it: "Is the number of inventory items for each content area representative of the amount of time typically devoted to each area in an introductory psychology course?" (1 = *not at all representative*, 5 = *very representative*), "How much of a need for this type of inventory is there in the field of psychology?" (1 = *no need at all*, 5 = *very strong need*), and "How likely are you to use the inventory in your course?" (1 = *not at all likely*, 5 = *very likely*).

We also asked participants, "If you were to use the inventory, how likely are you to use it for each of the following purposes?" (1 = *not at all likely*, 5 = *very likely*). The purposes rated were: "To assess how much learning is occurring in my course," "To assess the effect of new teaching techniques," "To fulfill institutional or departmental assessment requirements," "To compare different formats of the course (e.g., in-person vs. online format)," "To identify content areas students do not understand well," and "To have evidence of my teaching effectiveness." Finally, we asked demographic questions including institution type, number of times they had taught introductory psychology, their academic rank/title, and their area of specialty within psychology.

Results

Participants generally thought that the number of items for each content area was representative, with 97.4% indicating that the number of items per content area was "somewhat representative" or above. Furthermore, 89.5% of participants indicated there was "somewhat of a need" or more for this type of inventory in psychology, and 75.7% indicated they were at least "somewhat likely" to use the inventory in their course. Additionally, participants endorsed that they would use the inventory for a variety of purposes, with the most common purpose being to assess learning in their course (see Table 3).

Table 3
Survey Responses Regarding Inventory Uses

Item	Percent responding "somewhat likely" or above
If you were to use the inventory, how likely are you to use it for each of the following purposes? (1 = not at all likely, 5 = very likely)	
To assess how much learning is occurring in my course	97.4
To assess the effect of new teaching techniques	66.7
To fulfill institutional or departmental assessment requirements	86.8
To compare different formats of the course (e.g., in-person vs. online format)	70.3
To identify content areas students don't understand well	81.6
To have evidence of my teaching effectiveness.	76.3

We identified two items to remove from the inventory based on responses to the open-ended feedback question: one item from the Emotion area and one item from the Psychological Disorders and Therapy area (see supplemental materials). Six participants (15%) were concerned that the Psychological Disorders and Therapy item could have more than one correct answer. Three participants (7%) were concerned that the Emotion item was not representative of the intended theory. Although few participants indicated that the Emotion item was problematic, this item also showed poor discrimination in the item-level analyses for both Sample 1 and Sample 2 of Study 1 (see supplemental materials). Additionally, there was one other item (#20—Life Span Development) that three participants (7%) felt would be a "trick" question. However, we chose to keep this item in the inventory because the item-level analyses showed that it still contributed to discriminating low and high performers. Participants also identified other inventory items that they felt were not the best representation of a particular content area, or suggested creating new items for certain topics. However, except as already noted, fewer than three participants made any particular suggestion.

After removing the two items from the inventory, we reanalyzed the results of Study 1 and Study 2. There were only very small changes in the previously reported statistics (see supplemental materials), and therefore none of our previous conclusions changed.

Discussion

Results from Study 3 suggest that experienced introductory psychology faculty had positive feedback regarding the inventory. They largely thought the inventory content was representative of a typical introductory psychology course, that an inventory of this type is needed in the field of psychology, and that they would use the inventory to serve a variety of purposes including to assess learning in their course, fulfill institutional or departmental assessment requirements, and identify difficult content for students. After removing two items that received negative feedback, the inventory's validity and reliability remained robust. We take these data to mean that the inventory does assess knowledge that is representative of introductory psychology and the inventory would be of value to many faculty who routinely teach introductory psychology.

General Discussion

We sought to develop a reliable and valid assessment of introductory psychology knowledge. Using accepted test-development guidelines, we created a 32-item multiple-choice format inventory that spans 14 content areas of introductory psychology. Across two evaluation studies, we found evidence that the inventory has good test–retest reliability, and convergent, discriminant, and known-groups validity. After further feedback from a survey of introductory psychology instructors, we identified two items to remove. This resulted in a final 30-item inventory with comparable psychometric properties to the original inventory.[6]

A considerable strength of the current research was that we assessed the psychometric properties of the inventory in the population for which it is intended to be used: introductory psychology students. Specifically, we tested the inventory in six undergraduate introductory psychology courses in Study 1, which varied in their institution type, enrollment, and course

[6] The final inventory can be obtained by contacting the corresponding author.

type (i.e., in-person vs. online). The courses also varied in their instructor type, with the courses in Samples 1 and 2 having three instructors coteaching each course (resulting in six total instructors), and the courses in Samples 3–6 all having the same instructor. As such, these results have good external validity (i.e., generalizability).

In Study 3, we gathered feedback from a group of introductory psychology instructors, who indicated that the content of the inventory was representative, that an inventory of this type is needed, and that they would use the inventory in their courses. Specifically, faculty reported they would use the inventory for multiple purposes, including to assess student learning, to fulfill institutional or departmental assessment requirements, to identify content students do not understand well, to have evidence of their teaching effectiveness, to compare different formats of the course, and to assess the effects of new teaching techniques. With so many potential uses, it appears that the inventory could be valuable in many introductory psychology courses. Nevertheless, instructors should consider what purpose it will serve in their course or whether an existing assessment would better suit their needs. Furthermore, departments considering requiring the inventory to be administered to their introductory courses might consider whether mandatory use would cause difficulty for instructors (e.g., would they feel that they need to now "teach to the test") or impede their academic freedom.

One limitation of the current research is that the Cronbach's alpha reliability for the inventory was relatively low. Theoretically, this might reflect that the knowledge captured by this inventory is not unified by one common theme. This seems plausible considering the broad range of subdisciplines within psychology that are represented in the introductory course and on the inventory. Indeed, a student's knowledge of one content area on the inventory may differ considerably from their knowledge of another content area. However, we found evidence for good test–retest reliability, indicating that this inventory is likely a stable metric for assessing introductory psychology knowledge.

Practically speaking, this inventory is easy to administer and score, and takes only a short amount of time to complete. The ease of administration and scoring are aspects of the inventory that may promote adoption by instructors or department assessment committees. This may lead to more instructors and institutions conducting evaluations of their students' learning in introductory psychology, which would address the growing interest in university and program assessment (Dunn et al., 2007). In addition, the effectiveness of instructional techniques is essential knowledge for instructors. This inventory can be used to facilitate better understanding of the usefulness of various instructional techniques, and perhaps direct instructors toward evidence-based pedagogies.

While other measures of introductory psychology knowledge or general psychology knowledge exist (e.g., ETS, 2018a, 2018b; PACAT, Inc., 2018; Peter et al., 2015; Thompson & Zamboanga, 2004), none seemed like the ideal choice for assessing introductory psychology knowledge due to cost, lack of psychometric assessments, or ease of use. The inventory developed in the current research was developed specifically to address calls for reliable and valid assessments for the introductory course. As such, the content areas covered in the inventory align with the APA's five content pillars central to introductory psychology (APA, 2014). While this inventory does not cover content from every subfield within psychology, instructor feedback in Study 3 indicated that the content that is covered is sufficiently broad to be appropriate for many introductory courses. And although the majority of surveyed instructors indicated they were at least somewhat likely to use the inventory, instructors who do not cover all material contained in the inventory may find it less useful. A solution is to administer just the items that correspond to covered material.[7]

[7] We re-computed all analyses in Study 1 to determine whether the inventory was still reliable and valid when select sections were removed. Specifically, for Samples 1–2, we removed sections Emotion and Health and Well-Being because literature suggests these are the topics most likely not to be covered in the course (Bates, 2004). The results did not change when the analyses were re-computed without these two sections. Similarly, for Samples 3–6, we removed sections Sensation and Perception, Lifespan Development, Emotion, and Health and Wellbeing because they were not covered in the courses. Again, results did not change when the analyses were re-computed without these sections.

In addition to these strengths, some limitations of the inventory should be acknowledged. First, although other important learning outcomes have been identified for the introductory psychology course (e.g., skills; Jhangiani & Hardin, 2015; Strohmetz et al., 2015), the current inventory aimed to assess understanding of the content of the introductory psychology course. Second, although 97% of the instructors surveyed in Study 3 expressed interest in using the inventory to assess students' learning, we recognize that not all instructors may agree that the type of knowledge assessed in this inventory represents the type of learning that they value. Some may prefer an inventory that requires deeper conceptual analysis and excludes items that assess knowledge of terms that represent methods, theories, and so forth Third, it is certain that assessment of the introductory psychology course will evolve as the field of psychology continues to discuss the introductory course and the APA's working groups put forward their conclusions; thus, the current inventory may need to be updated over time. Nonetheless, at present the current inventory provides a reliable and valid tool for measuring learning in the introductory course.

References

Adams, W. K., & Wieman, C. E. (2011). Development and validation of instruments to measure learning of expert-like thinking. *International Journal of Science Education, 33,* 1289–1312. http://dx.doi.org/10.1080/09500693.2010.512369

American Psychological Association. (2013). *APA guidelines for the undergraduate psychology major: Version 2.0*. Retrieved from http://www.apa.org/ed/precollege/undergrad/index.aspx

American Psychological Association. (2014). *Strengthening the common core of the introductory psychology course*. Washington, DC: American Psychological Association, Board of Educational Affairs. Retrieved from http://www.apa.org/ed/governance/bea/intro-psych-report.pdf

American Psychological Association. (2018). *Board of Educational Affairs: Task forces and working groups reporting to BEA*. Retrieved from http://www.apa.org/ed/governance/bea/index.aspx

Bass, K. M., Drits-Esser, D., & Stark, L. A. (2016). A primer for developing measures of science content knowledge for small-scale research and instructional use. *CBE Life Sciences Education, 15,* 1–14. http://dx.doi.org/10.1187/cbe.15-07-0142

Bates, S. C. (2004). Coverage: Findings from a national sample of introductory psychology syllabi. In D. V. Doty (Chair), *What to leave in and out of introductory psychology*. Symposium conducted at the 112th convention of the American Psychological Association, Honolulu, HI.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. Lord & M. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397–479). Reading, MA: Addison Wesley.

Brownell, S. E., Hekmat-Scafe, D. S., Singla, V., Chandler Seawell, P., Conklin Imam, J. F., Eddy, S. L., . . . Cyert, M. S. (2015). A high-enrollment course-based undergraduate research experience improves student conceptions of scientific thinking and ability to interpret data. *CBE Life Sciences Education, 14,* 1–14. http://dx.doi.org/10.1187/cbe.14-05-0092

Butler, A. C., & Roediger, H. L., III. (2007). Testing improves long-term retention in a simulated classroom setting. *European Journal of Cognitive Psychology, 19,* 514–527. http://dx.doi.org/10.1080/09541440701326097

Cahill, M. J., Hynes, K. M., Trousil, R., Brooks, L. A., McDaniel, M. A., Repice, M., . . . Frey, R. F. (2014). Multiyear, multi-instructor evaluation of a large-class interactive-engagement curriculum. *Physical Review Special Topics Physics Education Research, 10,* 1–19. http://dx.doi.org/10.1103/PhysRevSTPER.10.020101

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56,* 81–105. http://dx.doi.org/10.1037/h0046016

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology, 78,* 98–104. http://dx.doi.org/10.1037/0021-9010.78.1.98

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16,* 297–334. http://dx.doi.org/10.1007/BF02310555

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52,* 281–302. http://dx.doi.org/10.1037/h0040957

Dorans, N. J. (1999). Correspondences between ACT™ and SAT® I scores. *ETS Research Report Series, 1999,* 1–18.

Dunn, D. S., Brewer, C. L., Cautin, R. L., Gurung, R. A. R., Keith, K. D., McGregor, L. N., . . . Voigt, M. J. (2010). The undergraduate psychology curriculum: Call for a core. In D. F. Halpern (Ed.), *Undergraduate education in psychology: A blueprint for the future of the discipline* (pp. 47–61). Washington, DC: American Psychological Association. http://dx.doi.org/10.1037/12063-003

Dunn, D. S., McCarthy, M. A., Baker, S., Halonen, J. S., & Hill, G. W., IV. (2007). Quality benchmarks in undergraduate psychology programs. *American Psychologist, 62,* 650–670. http://dx.doi.org/10.1037/0003-066X.62.7.650

Education Testing Service. (2018a). *ETS Major Field Test for Psychology*. Retrieved from https://www.ets.org/mft/about/content/psychology

Education Testing Service. (2018b). *GRE Psychology Test*. Retrieved from https://www.ets.org/gre/subject/about/content/psychology

Epstein, J. (2013). The Calculus Concept Inventory: Measurement of the effect of teaching methodology in mathematics. *Notices of the American Mathematical Society, 60,* 1018–1026. http://dx.doi.org/10.1090/noti1033

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39,* 175–191. http://dx.doi.org/10.3758/BF03193146

Garvin-Doxas, K., & Klymkowsky, M. W. (2008). Understanding randomness and its impact on student learning: Lessons learned from building the Biology Concept Inventory (BCI). *CBE-Life Sciences Education, 7,* 227–233. http://dx.doi.org/10.1187/cbe.07-08-0063

Geisinger, K. F. (2013). Reliability. In K. F. Geisinger (Ed.), *APA handbook of testing and assessment in psychology: Vol. 1. Test theory and testing and assessment in industrial and organizational psychology* (pp. 21–42). Washington, DC: American Psychological Association.

Griggs, R. A., Bujak-Johnson, A., & Proctor, D. L. (2004). Using common core vocabulary in text selection and teaching the introductory course. *Teaching of Psychology, 31,* 265–269.

Gurung, R. A., Hackathorn, J., Enns, C., Frantz, S., Cacioppo, J. T., Loop, T., & Freeman, J. E. (2016). Strengthening introductory psychology: A new model for teaching the introductory course. *American Psychologist, 71,* 112–124. http://dx.doi.org/10.1037/a0040012

Hake, R. R. (1998). Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses. *American Journal of Physics, 66,* 64–74. http://dx.doi.org/10.1119/1.18809

Hake, R. R. (2015). What might psychologists learn from Scholarship of Teaching and Learning in physics? *Scholarship of Teaching and Learning in Psychology, 1,* 100–106. http://dx.doi.org/10.1037/stl0000022

Hattie, J., & Cooksey, R. W. (1984). Procedures for assessing the validities of tests using the "known-groups" method. *Applied Psychological Measurement, 8,* 295–305. http://dx.doi.org/10.1177/014662168400800306

Haynes, S. N., Richard, D., & Kubany, E. S. (1995). Content validity in psychological assessment: A functional approach to concepts and methods. *Psychological Assessment, 7,* 238–247. http://dx.doi.org/10.1037/1040-3590.7.3.238

Henson, R. K. (2001). Understanding internal consistency reliability estimates: A conceptual primer on coefficient alpha. *Measurement and Evaluation in Counseling and Development, 34,* 177–189. http://dx.doi.org/10.1080/07481756.2002.12069034

Hestenes, D., Wells, M., & Swackhamer, G. (1992). Force concept inventory. *The Physics Teacher, 30,* 141–158. http://dx.doi.org/10.1119/1.2343497

Homa, N., Hackathorn, J., Brown, C. M., Garczynski, A., Solomon, E. D., Tennial, R., . . . Gurung, R. A. (2013). An analysis of learning objectives and content coverage in introductory psychology syllabi. *Teaching of Psychology, 40,* 169–174. http://dx.doi.org/10.1177/0098628313487456

Jhangiani, R. S., & Hardin, E. E. (2015). Skill development in introductory psychology. *Scholarship of Teaching and Learning in Psychology, 1,* 362–376. http://dx.doi.org/10.1037/stl0000049

Landrum, R. E., & Gurung, R. A. R. (2013). The memorability of introductory psychology revisited. *Teaching of Psychology, 40,* 222–227. http://dx.doi.org/10.1177/0098628313487417

Larsen, D. P., Butler, A. C., & Roediger, H. L., III. (2009). Repeated testing improves long-term retention relative to repeated study: A randomised controlled trial. *Medical Education, 43,* 1174–1181. http://dx.doi.org/10.1111/j.1365-2923.2009.03518.x

Libarkin, J. (2008, October). *Concept inventories in higher education science*. Retrieved from https://sites.nationalacademies.org/cs/groups/dbassesite/documents/webpage/dbasse_072624.pdf

McNamara, C. L., Williamson, A. L., & Jorgensen, T. D. (2011). Assessment of differential learning by topic in introductory psychology. *Psychology Learning & Teaching, 10,* 253–260. http://dx.doi.org/10.2304/plat.2011.10.3.253

Mitchell, K., Satterfield, J., Lewis, R. S., & Hong, B. A. (2016). The new Medical College Admission Test: Implications for teaching psychology. *The American Psychologist, 71,* 125–135. http://dx.doi.org/10.1037/a0039975

Nairn, S. L., Ellard, J. H., Scialfa, C. T., & Miller, C. D. (2003). At the core of introductory psychology: A content analysis. *Canadian Psychology/Psychologie canadienne, 44,* 93–99. http://dx.doi.org/10.1037/h0086930

Norcross, J. C., Hailstorks, R., Aiken, L. S., Pfund, R. A., Stamm, K. E., & Christidis, P. (2016). Undergraduate study in psychology: Curriculum

and assessment. *American Psychologist, 71,* 89–101. http://dx.doi.org/10.1037/a0040095

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York, NY: McGraw-Hill.

PACAT, Inc. (2018). *ACAT online: Psychology*. Retrieved from https://www.collegeoutcomes.com/NLI/dsp/dsp_03.aspx

Peck, A. C., Ali, R. S., Levine, M. E., & Matchock, R. L. (2006). Introductory psychology topics and student performance: Where's the challenge? *Teaching of Psychology, 33,* 167–170. http://dx.doi.org/10.1207/s15328023top3303_2

Peter, J., Leichner, N., Mayer, A. K., & Krampen, G. (2015). A short test for the assessment of basic knowledge in psychology. *Psychology Learning & Teaching, 14,* 224–235. http://dx.doi.org/10.1177/1475725715605763

Proctor, D. L., & Williams, A. M. (2006). Frequently cited concepts in current introduction to psychology textbooks. *Society for Teaching of Psychology Office of Teaching Resources in Psychology*. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.690.7972&rep=rep1&type=pdf

Rizopoulos, D. (2006). ltm: An R package for latent variable modeling and item response theory analyses. *Journal of Statistical Software, 17,* 1–25. http://dx.doi.org/10.18637/jss.v017.i05

Smith, M. K., Wood, W. B., & Knight, J. K. (2008). The genetics concept assessment: A new concept inventory for gauging student understanding of genetics. *CBE-Life Sciences Education, 7,* 422–430. http://dx.doi.org/10.1187/cbe.08-08-0045

Strohmetz, D. B., Dolinsky, B., Jhangiani, R. S., Posey, D. C., Hardin, E. E., Shyu, V., & Klein, E. (2015). The skillful major: Psychology curricula in the 21st century. *Scholarship of Teaching and Learning in Psychology, 1,* 200–207. http://dx.doi.org/10.1037/stl0000037

Thompson, R. A., & Zamboanga, B. L. (2004). Academic aptitude and prior knowledge as predictors of student achievement in introduction to psychology. *Journal of Educational Psychology, 96,* 778–784. http://dx.doi.org/10.1037/0022-0663.96.4.778

Veilleux, J. C., & Chapman, K. M. (2017a). Development of a research methods and statistics concept inventory. *Teaching of Psychology, 44,* 203–211. http://dx.doi.org/10.1177/0098628317711287

Veilleux, J. C., & Chapman, K. M. (2017b). Validation of the Psychological Research Inventory of Concepts: An index of research and statistical literacy. *Teaching of Psychology, 44,* 212–221. http://dx.doi.org/10.1177/0098628317711302

Zechmeister, J. S., & Zechmeister, E. B. (2000). Introductory textbooks and psychology's core concepts. *Teaching of Psychology, 27,* 6–11. http://dx.doi.org/10.1207/S15328023TOP2701_1

Received June 10, 2019
Revision received August 25, 2019
Accepted August 26, 2019 ■