Department of Planning, Innovation, and Accountability

Research Brief

Report from the Office of Student Assessment February 10, 2015

Validation Study of the Integrated Performance Task

AUTHOR: Douglas G. Wren, Ed.D., Assessment Specialist
Department of Planning, Innovation, and Accountability

OTHER CONTACT PERSON: Donald E. Robertson, Jr., Ph.D., Assistant Superintendent
Department of Planning, Innovation, and Accountability

ABSTRACT

During spring 2013 and spring 2014, data for a two-phase study designed to obtain criterion-related validity evidence for the Integrated Performance Task (IPT) were collected at ten elementary and three middle schools in Virginia Beach City Public Schools (VBCPS). The IPT is a measure of critical thinking, problem solving, and written communication administered twice annually to VBCPS students in grades 4 and 7. In addition to the IPT, the California Critical Thinking Skills Test (CCTST), a multiple-choice test of critical-thinking skills, was administered to students in grades 4 and 7. The results were correlated to determine if any significant relationships existed between the attributes purported to be measured by the assessments. Moderate correlations between IPT scores and the overall scores on the CCTST provided evidence that IPT scores are valid indicators of critical-thinking skills for students in grade 4 and grade 7.

"Validation is the process of examining the accuracy of a specific prediction or inference made from a test score."[1] —Lee J. Cronbach (1971)

BACKGROUND

This study compared the scores of Virginia Beach City Public Schools (VBCPS) students in grades 4 and 7 on the Integrated Performance Task (IPT) with their scores on the California Critical Thinking Skills Test (CCTST). The purpose of the study was to analyze the relationship between IPT and CCTST scores to gain a better understanding of what the IPT is measuring among students in the targeted grades.

Psychometricians—professionals with expertise and training in educational and psychological measurement—would describe this research as a validation study. According to the *Standards for Educational and Psychological Testing*, validity is "the most fundamental consideration in developing tests

Key Topics:

[1]Cronbach, L. J. (1971). Test Validation. In R. L. Thorndike (Ed.), *Educational Measurement* (2nd ed.). Washington, DC: American Council on Education.

and evaluating tests."[2] Evidence of validity for a test is essential in order for assumptions to be made about what the test is actually measuring. For example, a mathematics test comprised of word problems may be intended to assess students' proficiency in using mathematical operations, but if the problems incorporate complex wording, the test could actually be measuring reading comprehension. Issues such as readability should be addressed during the test development process; it should also be determined if the test content is age-appropriate and representative of the domain that the test is supposed to measure (e.g., multiplication, adverbs, U.S. history).

Messick suggested the following procedure for obtaining content-related evidence of validity: "Appraise the relevance and representativeness of the test content in relation to the content of the behavioral or performance domain about which inferences are to be drawn or predictions made."[3] Typically, content-related validity evidence is acquired through expert judgment. In other words, qualified persons with testing and test content expertise appraise the relevance of the assessment and decide if it is representative of the performance domain.

Each IPT was developed by a team of VBCPS educators with decades of collective experience in teaching and testing. These experts worked and reviewed these performance tasks to ensure that they were grade-level appropriate and representative of the performance domains for which they were intended. Subsequently, the IPTs were evaluated by other education and testing professionals. The compilation of these judgments provided content-related evidence of validity for the IPT. It should be noted that this type of evidence is insufficient by itself because it is limited to the adequacy and representativeness of the assessment for testing purposes only.

A criterion-related validation study is a method of obtaining validity evidence by checking test scores against external criteria. Using this method, the results of a relatively new test—such as the IPT—are correlated with an established test, or criterion, that measures the same psychological construct. A construct is defined as a "postulated attribute of people, assumed to be reflected in test performance."[4] Some examples of constructs are anxiety, self-esteem, and intelligence. The primary construct of interest in this study was critical thinking, one of the outcomes for student success designated by *Compass to 2015*, the strategic plan adopted in 2008 by the Virginia Beach School Board. The criterion tests for this study were the elementary and middle school versions of the CCTST. The IPT and CCTST are described in the next section.

MEASURES

The Integrated Performance Task

The IPT is the acronym given to performance tasks that were developed by VBCPS staff. These assessments have been administered to students in grades 4 and 7 since October 2010. The first versions of the IPT were created during the 2009-2010 school year under the direction of a subcommittee comprised of select teachers, staff from the Department of Teaching and Learning, and an assessment specialist from the Department of Planning, Innovation, and Accountability. The subcommittee was commissioned by the *Compass to 2015* Strategic Objective 2 Action Team and called the "CWRA-Type Assessment Development Team" because the IPT was modeled after the College and Work Readiness Assessment (CWRA).

---

[2]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.

[3]Messick, S. (1990). *Validity of Test Interpretation and Use*. Princeton, NJ: Educational Testing Service, (ERIC Document Reproduction Service No. ED395031) <http://files.eric.ed.gov/fulltext/ED395031.pdf>, accessed on December 15, 2014.

[4]Cronbach, L. J., & Meehl, P.E. (1955). Construct Validity in Psychological Tests. *Psychological Bulletin, 52*.

The CWRA is an online performance task that assesses high school students' readiness for college or work by evaluating their skills in analytic reasoning and evaluation, problem solving, writing effectiveness, and writing mechanics. Each CWRA performance task presents students with an engaging, real-world scenario and a set of accompanying documents. After reviewing the scenario and documents, students must construct responses to explain their recommendations for solving the problem embedded within the performance task. Like the CWRA, each IPT provides a realistic scenario and open-ended questions tailored for either fourth- or seventh-grade students. Unlike the CWRA, which is administered completely online, the IPT scenario and documents are presented to students in printed booklets. After reviewing the booklet, students are required to type their responses to the questions, called prompts, on a laptop or desktop computer and submit their responses on Schoolnet, the online test administration system used by VBCPS.

At the time that data were collected, there were three prompts on both the grade 4 and the grade 7 IPT. For the three-prompt IPT, students' responses were scored in two areas of critical thinking (CT1 and CT2), and one each for problem solving (PS) and written communication (WC). The areas are called elements on the IPT rubric, which provides the basis for scoring IPT responses. Each response is scored on a scale from 1 to 4, with the four scoring levels designated as follows: Level 1 = Novice, Level 2 = Emerging, Level 3 = Proficient, Level 4 = Advanced. Table 1 shows the rubric elements and each element's alignment with the prompts for the IPTs that were administered for this study.

Table 1
IPT Rubric Elements and Prompts

Element	Definition	Aligned With	Description of Prompt
CT1	Decides if the information in the IPT booklet is correct and believable.	Prompt 1	Find examples of incorrect, unbelievable, or misleading information in a specific document and explain why the information is incorrect, unbelievable, or misleading.
CT2	Sees the need for important information not in the IPT booklet.	Prompt 2	Describe relevant information that was not in the booklet or not clearly explained and tell why the missing or incomplete information is needed to help make a wise choice.
PS	Makes a choice and gives reasons for the choice.	Prompt 3	Write a letter or a recommendation to explain the choice and support the recommendation with relevant reasons.
WC	Presents information and ideas that are clear, organized, detailed, and written for the intended audience.	Prompt 3	Include a greeting, an opening sentence that states the recommendation or choice, reasons that support the recommendation or choice, a final sentence that summarizes the recommendation or choice, and a closing.

The IPT is generally administered to an entire class of students by their teacher. As the students follow along in their booklets, grade 4 teachers read the situation, the documents, and the prompts aloud. Grade 7 teachers read only the situation aloud to their students. However, any student in either grade may ask to have a word, phrase, or sentence reread or explained to them. So teachers do not provide their own subjective definitions to students, a glossary of terms is included in the grade 4 teacher directions and at the back of the grade 7 student booklet. Students are given 90 minutes to complete the IPT after the teacher finishes reading all of the directions. Students with documented testing accommodations on an Individualized Education Program (IEP) or a 504 Plan can be given the IPT individually or in a small group.

The California Critical Thinking Skills Test

The CCTST family is a set of critical thinking skills tests marketed by Insight Assessment, a division of California Academic Press. All of the tests are based on the definition of critical thinking that emerged from a two-year Delphi study conducted by the American Philosophical Association. The consensus definition stated that critical thinking is "purposeful, self-regulatory judgment which results in interpretation, analysis, evaluation, and inference, as well as explanation

of the evidential, conceptual, methodological, criteriological, or contextual considerations upon which that judgment is based."[5] Different versions of the CCTST have been developed to measure the critical-thinking skills of students in kindergarten through graduate school and adults in various professions, including military and defense personnel. The skills for the CCTST-MIB and CCTST-M25—designed for students in grades 3-5 and 6-9, respectively—are described in Table 2.

Table 2
Skills Measured by the CCTST

Skill	Description[6]
Overall	The overall score describes overall strength in using reasoning to form reflective judgments about what to believe or what to do. To score well overall, the test-taker must excel in the sustained, focused, and integrated application of core reasoning skills including analysis, interpretation, inference, evaluation, explanation, induction, and deduction. The overall score predicts the capacity for success in educational or workplace settings which demand reasoned decision making and thoughtful problem solving.
Induction	Decision making in contexts of uncertainty relies on inductive reasoning. We use inductive reasoning skills when we draw inferences about what we think must probably be true based on analogies, case studies, prior experience, statistical analyses, simulations, hypotheticals, and familiar circumstances, and patterns of behavior. As long as there is the possibility, however remote, that a highly probable conclusion might be mistaken, the reasoning is inductive. Although it does not yield certainty, inductive reasoning can provide a solid basis for confidence in our conclusions.
Deduction	Decision making in precisely defined contexts where rules, operating conditions, core beliefs, values, policies, principles, procedures, and terminology completely determine the outcome depends on strong deductive reasoning skills. Deductive reasoning moves with exacting precision from the assumed truth of a set of beliefs to a conclusion which cannot be false if those beliefs are true. Deductive validity is rigorously logical and clear-cut. Deductive validity leaves no room for uncertainty, unless one alters the meanings of words or the grammar of the language.
Analysis	Analytical reasoning skills enable people to identify assumptions, reasons and claims, and to examine how they interact in the formation of arguments. We use analysis to gather information from charts, graphs, diagrams, spoken language, and documents. People with strong analytical skills attend to patterns and to details. They identify the elements of a situation and determine how those elements interact. Strong interpretation skills can support high-quality analysis by providing insights into the significance of what a person is saying or what something means.
Inference	Inference skills enable us to draw conclusions from reasons and evidence. We use inference when we offer thoughtful suggestions and hypotheses. Inference skills indicate the necessary or the very probable consequences of a given set of facts and conditions. Conclusions, hypotheses, recommendations, or decisions that are based on faulty analyses, misinformation, bad data, or biased evaluations can turn out to be mistaken, even if they have been reached using excellent inference skills.
Evaluation	Evaluative reasoning skills enable us to assess the credibility of sources of information and the claims they make. We use these skills to determine the strength or weakness of arguments. Applying evaluation skills, we can judge the quality of analyses, interpretations, explanations, inferences, options, opinions, beliefs, ideas, proposals, and decisions. Strong explanation skills can support high-quality evaluation by providing the evidence, reasons, methods, criteria, or assumptions behind the claims made and the conclusions reached.
Numeracy	Numeracy skills are used when applying knowledge of numbers, arithmetic, measures, and mathematical techniques to situations that require the interpretation or evaluation of information. Numeracy refers to the ability to solve quantitative reasoning problems, or to make judgments derived from quantitative reasoning in a variety of contexts. More than being able to compute a solution to a mathematical equation, numeracy includes the understanding of how quantitative information is gathered, manipulated, and represented visually, such as in graphs, charts, tables, and diagrams.

[5]Facione, P. A. (1990). *Critical Thinking: A Statement of Expert Consensus for Purposes of Educational Assessment and Instruction.* Newark, DE: American Philosophical Association. (ERIC Document Reproduction Service No. ED315423) <http://files.eric.ed.gov/fulltext/ED315423.pdf>, accessed on December 15, 2014.

[6]Insight Assessment. (2013). *California Critical Thinking Skills Test: CCTST Test Manual.* San Jose, CA: California Academic Press.

The MIB and M25 versions of the CCTST both consist of multiple-choice items administered online or in paper/pencil format. There are 20 items on the CCTST-MIB and 25 items on the CCTST-M25. Most of the items on the assessments are word problems, and several items include a related chart, graph, or image for test takers to view in order to help them choose the correct answer among four response options on the CCTST-MIB and five options on the CCTST-M25. Students are given up to 45 minutes to complete the test. For each student, an overall score and subscores for each of the six skills described in Table 2 are reported as standard scores with a range of 60 to 100. In addition, individual student reports provide four corresponding designations for the scores, listed from lowest to highest: Not Manifested, Emerging, Strong, and Superior.

The CCTST was selected as the criterion measure for this study for several reasons. Crocker and Algina stated that "there must be a judicious tradeoff in selecting a criterion which (1) can be reliably measured within the time and cost constraints of the study and (2) will have a relationship to the ultimate criterion of interest to most test users."[7] In this case, the ultimate criterion of interest was critical thinking. The CCTST-MIB and CCTST-M25 are easy to administer, well within the time and cost limits, and the items are appropriate for fourth- and seventh-grade students.

The CCTST has established evidence of validity and satisfactory reliability. According to the CCTST M-Series Test Manual, evidence of content-related validity was acquired "with the assistance of educators in schools, child development programs and educational settings throughout the USA."[8] Criterion-related evidence of validity for the CCTST M-Series continues to accumulate through studies conducted by doctoral students and other independent researchers. Using Kuder-Richardson 20, the internal consistency estimates of the CCTST-M series range from .78 to .82. These values indicate that all of the items on the CCTST-MIB and CCTST-M25 adequately contribute to the measurement of a single attribute or construct—critical thinking.

PARTICIPANTS

Grade 4

The students who participated in the first phase of this study were enrolled in fourth grade at a VBCPS elementary school during the 2012-2013 school year. Principals at 10 of 54 schools that house fourth-grade classes expressed their interest in participating by responding to a memo inviting schools to take part in the study. One fourth-grade class at each of the ten schools took the CCTST-MIB during the same test administration window that was scheduled for the IPT: March 25 through April 19, 2013. The scores for the CCTST-MIB were accessible immediately because the assessment was administered online, while the results of the IPT were not available until all of the spring IPT responses were scored in July 2013. Results for both assessments were received for a total of 207 students, although a few students did not respond to all three prompts on the IPT.

The elementary schools involved in the first phase of the study were located in diverse geographical sections of Virginia Beach. In all, seven different zip code zones were represented. The classes included three types of instructional settings: general education classes, inclusion classes, and gifted cluster classes. General education classes consist mainly of students who have not been identified as gifted and are not "students with disabilities" (i.e., SWD). Inclusion classes contain SWD and general education students, and gifted cluster classes include general education students and "a group (cluster) of identified gifted students… assigned to a classroom with a cluster

[7]Crocker, L., & Algina, J. (1986). Introduction to Classical and Modern Test Theory. New York, NY: Holt, Rinehart and Winston.

[8]Insight Assessment. (2013). *California Critical Thinking Skills Test: CCTST Test Manual.* San Jose, CA: California Academic Press.

teacher who collaborates with the gifted resource teacher to provide differentiated curriculum and instruction."[9]

Figure 1 below shows the number of grade 4 students who participated in the study by school zip code as well as the percentages of students by type of instructional setting within each zip code. Five grade 4 gifted cluster classes participated in the study. There were also three general education and two inclusion classes that took part in the first phase. Exactly 28 percent of the fourth-grade students who participated in the study attended the three schools located in the 23462 zip code zone. About 22 percent of the fourth-grade participants were enrolled in gifted cluster classes in the two schools located in the 23454 zone. The other zip code zones—23451, 23452, 23455, 23457, and 23464—each comprised from 7.7 to 12.1 percent of the students who participated in the first phase of the study.

Although nearly 53 percent of the participants in the study's first phase were enrolled in a gifted cluster class, there were only 35 identified gifted students out of the 109 children in these classes. This was approximately 17 percent of the 207 fourth-grade participants. Because general education students made up the majority of students in every inclusion and gifted cluster class in the first phase of the study, it was determined that well over 75 percent of the participants in the fourth-grade sample were general education students.

Figure 1
Instructional Settings and School Zip Codes of Grade 4 Participants (n = 207)

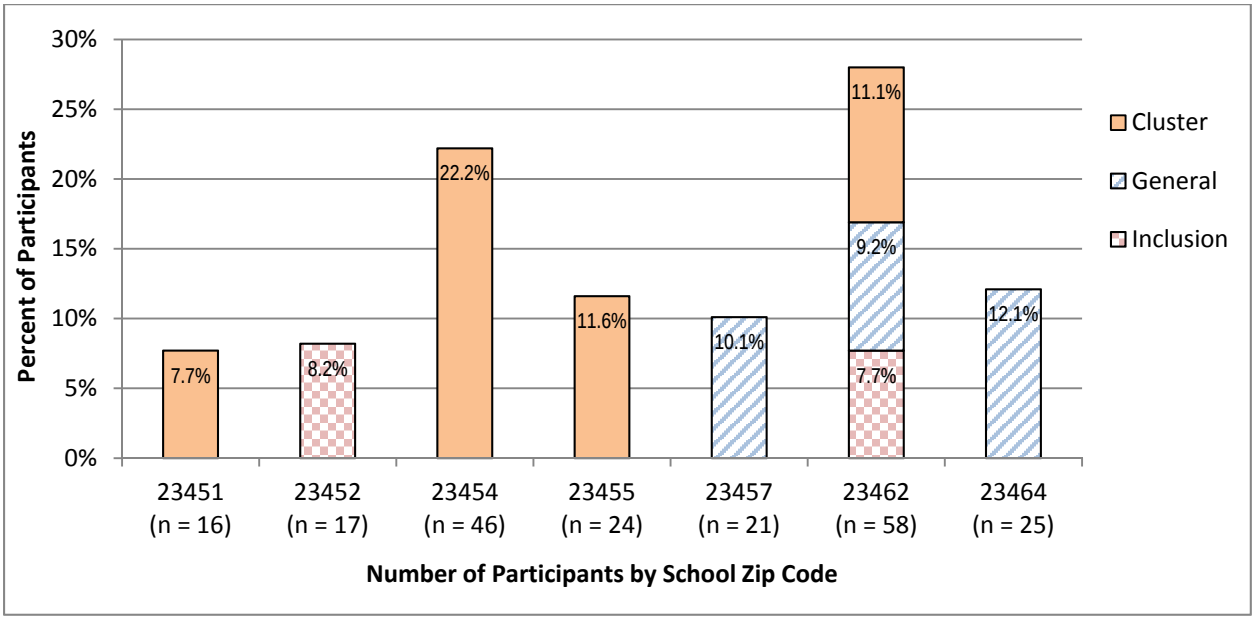


Table 3 on the next page provides the number and percentages of students by gender and ethnic group for the fourth-grade sample. For comparison purposes, the group percentages for all VBCPS students enrolled in grade 4 during the testing window are also included in Table 3. In general, the sample reflected the population of fourth-grade students in Virginia Beach in terms of gender and ethnicity at the time the assessments were administered. There was a difference of less than 4 percentage points between each group represented in the sample and the corresponding group in the VBCPS population. For the multiracial (i.e., two or more ethnicities) and African American groups, the difference between the sample and population was 1 point or less.

[9]Virginia Beach City Public Schools, Office of Gifted Education and Curriculum Development. (2012). *Elementary Gifted Resource Program Resource-Cluster Model*. Virginia Beach, VA: Virginia Beach City Public Schools.

Table 3
Demographic Summary of Grade 4 Validation Study Participants (n = 207)

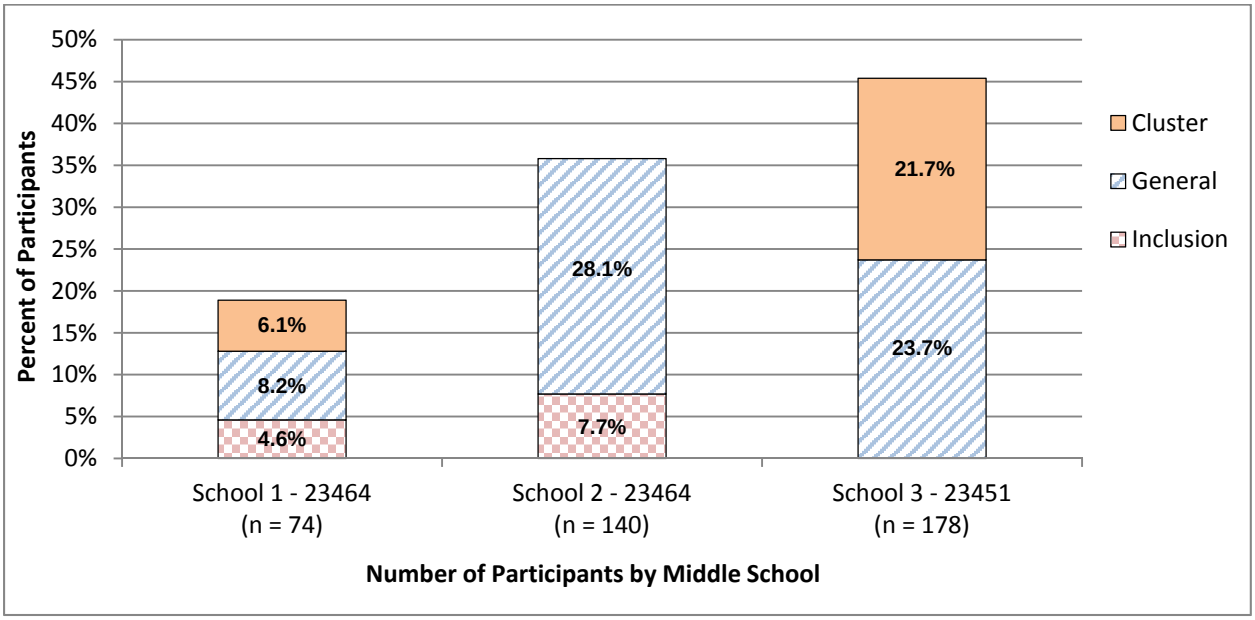
Characteristic	Number	Percent Sample	Percent VBCPS
Gender			
Female	110	53.1%	49.2%
Male	97	46.9%	50.8%
Ethnicity			
African American/Black	49	23.7%	23.2%
Asian	6	2.9%	5.6%
Hispanic	17	8.2%	9.5%
Two or More Ethnicities	19	9.2%	8.2%
White	116	56.0%	52.5%

Grade 7

The second phase of the study involved students enrolled in grade 7 during the 2013-2014 school year. The procedure for recruiting classes was similar to the first phase; all VBCPS middle school principals received a memo inviting their schools to take part in the study. Three principals agreed to have several grade 7 classes at their schools each take the CCTST-M25 during the IPT administration window of March 24 through April 11, 2014. Once again, the results for the online multiple-choice test were accessible immediately, while the IPT scores became available after all of the spring IPT responses were scored in July 2014. Usable results for both tests were obtained for 392 seventh-grade students, even though a few students did not respond to each prompt.

The middle schools that participated in this phase were located in two zip code zones of Virginia Beach. Figure 2 shows the number of students at each school and the percent of students in the entire sample by class type (i.e., gifted cluster, general education, or inclusion). Participants in School 1 and School 2 in the 23464 zip code zone combined to make up over 54 percent of participants in seventh grade. The remainder were students at School 3, located in the 23451 zone.

Figure 2
Student Classifications and School Zip Codes of Grade 7 Participants (n = 392)



In all, 14 seventh-grade classes participated in the second phase: three classes at School 1, five classes at School 2, and six classes at School 3. As Figure 2 illustrates, a majority of the participants—60 percent—were housed in the eight general education classes that took part in the study. Nearly 28 percent of the participants were in four gifted cluster classes, and the remaining participants were in the two inclusion classes in the 23464 zip code zone. The School 3 sample did not include any inclusion classes, although three School 3 participants were classified as SWD. There were no participating gifted cluster classes from School 2, but a number of students identified as gifted took part in the study at School 2. Many of the students in the gifted cluster and inclusion classes were general education students. Consequently, general education students comprised over 76 percent of the sample, with a large proportion of these students enrolled at School 3.

The number and percentages of students by gender and ethnic group for the grade 7 sample as well as for the VBCPS grade 7 population during the testing window are shown in Table 4. The difference between the group percentages for the sample and the population ranged from to 0.5 to 6.6 points. It was therefore ascertained that the sample adequately represented the population of grade 7 students in Virginia Beach in terms of gender and ethnicity for the purposes of the study.

Table 4
Demographic Summary of Grade 7 Validation Study Participants (n = 392)

Characteristic	Number	Percent Sample	Percent VBCPS
Gender			
Female	197	53.1%	49.4%
Male	195	46.9%	50.6%
Ethnicity			
African American/Black	73	18.6%	25.2%
American Indian	3	0.8%	0.3%
Asian	27	6.9%	5.5%
Hispanic	29	7.4%	9.6%
Two or More Ethnicities	39	10.0%	8.0%
White	221	56.4%	50.8%

METHOD

As data for each phase of the study became available, the results were correlated to obtain statistical interpretations of the relationship between IPT and CCTST scores for the samples. Correlation values range from –1 to +1, which indicate the direction and strength of relationships. When the scores of both measures for participants in a sample tend to increase, there will be a positive correlation between the measures. For example, a positive correlation value would be expected in a study comparing high school students' grades in math courses and their scores on the SAT mathematics section. Negative correlation values occur when the scores for one measure increase while the scores on the other measure decrease. A negative correlation would be likely in a study that compares grade point averages of students with the number of hours they play video games each day.

For this study, nonparametric correlations were run using IBM SPSS Statistics 20.[10] Nonparametric correlations are recommended for ordinal measures and the IPT yields ordinal data

[10]The acronym SPSS originally stood for Statistical Package for the Social Sciences. SPSS is a software package used for statistical analyses. <http://www.spss.ie/software/software_faq.html>, accessed on December 17, 2014.

(i.e., scores that are rank ordered). The analyses yielded 28 values for correlations between the four IPT elements and the seven CCTST scores for the fourth-grade sample and another 28 values for the seventh-grade sample. Each correlation was tested for statistical significance—an indication of the probability that the relationship did not occur by chance—and the correlations were also examined to determine the strength of the relationships between the different IPT elements and CCTST skills. These correlations are known as validity coefficients.

There is no universal agreement concerning acceptable values for validity coefficients in research. According to Cronbach "it is unusual for a validity coefficient to rise above 0.60."[11] Cronbach further stated that the only sensible answer to the question, "What is a good validity coefficient?" is, "The best you can get." Regardless, there are rules of thumb for interpreting validity coefficients. One source noted that tests with validity coefficients ranging from .21 to .35 are "likely to be useful," while tests with coefficients above .35 are "very beneficial."[12] Cohen designated correlations of .10 as small, .30 as moderate, and .50 as large.[13] For the current study, the predetermined cutoff was .30; validity coefficients of .30 or higher would indicate that an IPT element is measuring, to a moderate degree, the same attribute as a skill area on the CCTST.

Another question that typically arises in research such as in the present study is whether the sample size is sufficient to detect statistically significant relationships. A relationship that has a moderate correlation within a population may not be identified as statistically significant with an inadequate sample size. Some researchers have suggested that samples of 200 or more participants are needed to accurately reflect validity levels of populations at least 90 percent of the time.[14] Given that schools were not required to participate in this study, it was fortunate that both samples exceeded 200 students.

RESULTS

Tables 5 and 6 on the next page show the correlations or validity coefficients for the students tested in the first phase and second phase, respectively. As mentioned previously, not all students responded to all three prompts on the IPT. For fourth grade, one student did not answer Prompt 1, four students did not respond to Prompt 2, and three students did not write a response to Prompt 3. During the second phase, two seventh-grade students did not respond to Prompt 1, another student did not answer Prompt 2, and one other student did not provide a response to Prompt 3. These students' scores were not included in the analyses for the missing elements.

The strongest correlations for the fourth-grade sample were for the following pairs of IPT elements and CCTST-MIB scores: WC and Induction (.37), CT1 and Overall (.36), CT1 and Induction (.35), and WC and Overall (.35). The IPT element of PS had moderate correlations with Numeracy (.34), Overall (.33), Induction (.33), and Analysis (.33). Other moderate correlations were observed for CT1 and Inference (.33), WC and Evaluation (.31), WC and Numeracy (.31), and CT1 and Evaluation (.31). Of the four IPT elements, CT2 yielded the lowest correlations. The CCTST-MIB Deduction subscore had low correlations with each IPT element, ranging from .17

---

[11]Cronbach. L. J. (1970). *Essentials of Psychological Testing* (3rd ed.). New York, NY: Harper and Row.

[12]U.S. Department of U.S. Department of Labor, Employment and Training Administration. (1999). *Testing and Assessment: An Employer's Guide to Good Practices.* <http://uniformguidelines.com/testassess.pdf>, accessed on December 17, 2014.

[13]Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

[14]Schmidt, F. L., Hunter, J. E., & Urry, V. W. (1976). Statistical Power in Criterion-Related Validation Studies. *Journal of Applied Psychology,* 61.

to .20. Not surprisingly, the correlation between Deduction and CT2 was the lowest value for the grade 4 samples in the first and second phases of the study.

Table 5
Correlations Between the Grade 4 IPT and CCTST-MIB Scores (n = 207)

CCTST Skills	IPT Elements			
	Critical Thinking 1 (n = 206)	Critical Thinking 2 (n = 203)	Problem Solving (n = 204)	Written Communication (n = 204)
Overall	**.36***	.29*	**.33***	**.35***
Induction	**.35***	.27*	**.33***	**.37***
Deduction	.18**	.17**	.20*	.18*
Analysis	.26*	.21*	**.33***	.29*
Inference	**.33***	.29*	.25*	.29*
Evaluation	**.31***	.26*	.29*	**.31***
Numeracy	.23*	.19*	**.34***	**.31***

Note. Values of .30 or greater are in bold and indicate moderate correlations.
*Correlation is significant at the .01 level of probability (2-tailed).
**Correlation is significant at the .05 level of probability (2-tailed).

As shown in Table 6 below, correlations of .40 or greater were obtained for the seventh-grade sample for six pairs of IPT elements and CCTST-MIB scores: WC and Overall (.42), WC and Induction (.42), CT2 and Overall (.41), WC and Evaluation (.40), CT2 and Evaluation (.40), and CT2 and Analysis (.40). Additional moderate correlations ranging from .30 to .39 were observed for 16 other pairs of scores, including CT1 and Overall (.37). On average, WC had the highest grade 7 IPT correlations with CCTST scores, followed closely by CT2. The lowest correlations on the grade 7 IPT were for the PS element. The implications of the results for the study will be addressed in the following section.

Table 6
Correlations Between the Grade 7 IPT and CCTST-M25 Scores (n = 392)

CCTST Skills	IPT Elements			
	Critical Thinking 1 (n = 390)	Critical Thinking 2 (n = 391)	Problem Solving (n = 391)	Written Communication (n = 391)
Overall	**.37***	**.41***	**.31***	**.42***
Induction	**.35***	**.39***	**.30***	**.42***
Deduction	**.30***	**.35***	.25*	**.34***
Analysis	**.37***	**.40***	.27*	**.38***
Inference	**.34***	**.31***	.27*	**.35***
Evaluation	**.31***	**.40***	.29*	**.40***
Numeracy	.29*	**.35***	.23*	**.34***

Note. Values of .30 or greater are in bold and indicate moderate correlations.
*Correlation is significant at the .01 level of probability (2-tailed).

DISCUSSION

All 56 validity coefficients—correlations between scores for all seven CCTST skills and the scores for all four elements of the IPT for both samples—obtained in this study were statistically significant. This was due to the tendency for participants in the study with high scores on the IPT also scoring high on the CCTST, participants with average IPT scores performing at an average level on the CCTST, and participants with low scores on the IPT scoring low on the CCTST. All but two validity coefficients were significant at the .01 level of probability, which implies that the likelihood of these relationships occurring by chance is less than one percent. That is, the certainty is quite good that the CCTST and the IPT are measuring, to some degree, the same constructs.

The primary construct of interest in this study was critical thinking. The validity coefficients for the overall score on the CCTST and the CT1 element on the IPT were .36 and .37 for the grade 4 and grade 7 samples, respectively. This suggests CT1 is measuring to a moderate degree similar constructs that the CCTST is measuring. Further, these correlations suggest that the ability to identify incorrect, unbelievable, or misleading information and explain why the information is not credible may be an indication of critical thinking at the upper elementary and middle school levels.

There were other validity coefficients worth noting. The strongest correlations besides CT1 and Overall for the fourth-grade sample were between WC and Induction (.37) and WC and Overall (.35). These two pairs also had the highest correlations (both at .42) for the seventh-grade sample. The grade 7 correlations between WC and the other five CCTST subscores ranged from .34 to .40, while the same correlations for grade 4 were lower, ranging from .18 to .31. The relationship between writing ability and critical thinking will be addressed later in this section.

The validity coefficients for CT1 and Induction were .35 for both samples. These correlations indicated that students who performed well on the IPT demonstrated strong inductive reasoning skills on the CCTST (and vice versa) and suggest that CT1 is a modest measure of induction as operationalized by the CCTST. Because of the relationship between the concepts of induction and inference—a discussion of their similarities and differences extends well beyond the scope of this brief—it was not unexpected that there were moderate correlations between CT1 and Inference for the fourth-grade (.33) and seventh-grade (.34) samples.

It was interesting to note the grade level differences of several validity coefficients involving Analysis, Numeracy, CT1, and PS. For the grade 4 sample, moderate correlations were observed between PS and Analysis (.33) and PS and Numeracy (.34). The correlations for the grade 7 sample were much lower: PS and Analysis (.27), and PS and Numeracy (.23). Conversely, the value of the seventh-grade correlation between CT1 and Analysis (.37) was much higher than it was for the fourth-grade correlation between CT1 and Analysis (.26). While it might be assumed that good problem solvers tend to have effective analytical and quantitative reasoning skills, this was not the case for the older sample. Analytical thinking was related more to the ability to identify and explain examples of incredible information for seventh-grade students than for fourth-grade students. These inconsistencies may have been due to possible developmental differences between fourth and seventh graders, disparities between the two forms of both assessments, unknown variables, or a combination of these reasons.

Two questions emerged regarding the correlations generated from the data obtained in this study. The first question is why several of the strongest correlations for both samples involved written communication. From the inception of the IPT, the developers realized that there would be some overlap between the elements being measured. For instance, a person's critical-thinking skills would naturally be used in many problem-solving situations. Likewise, a cogent description of a solution to a problem requires sound written communication skills on the part of the problem solver.

In order to investigate these assumptions, correlations were run between the elements of the IPT for both samples (see the tables in the appendix). Each correlation was statistically significant at the .01 level of probability. The correlations between WC and PS had the highest values for fourth grade (.59) and for seventh grade (.56). This suggests that problem-solving skills are strongly related to writing ability on the IPT. For the fourth-grade sample, the next highest value was for WC and CT1 (.37), and for the seventh-grade sample, the second-highest value involved WC and CT2 (.44). The latter value was much higher than it was for WC and CT2 (.29) for the fourth-grade sample; discrepancies between the two samples for the CT2 element is the other question to be addressed.

The absence of validity coefficients at .30 or higher for CT2 for the fourth-grade sample can be easily explained. The reason for values ranging from .31 to .41 for all of the seventh-grade CT2 validity coefficients can be surmised. Beginning with the first administration of the spring IPT, it was noticeable that most fourth graders did not understand or were not adequately prepared to respond to Prompt 2, aligned with CT2. (Prompt 2 is worded, "Describe relevant information that was not in the booklet or not clearly explained and tell why the missing or incomplete information is needed to help make a wise choice.") Well over half of the responses to Prompt 2 were scored at Level 1 for the grade 4 spring IPT administrations in 2011, 2012, and 2013. Furthermore, a number of fourth-grader students over the years did not respond to this prompt and were given a score of 0. It can be assumed that the sample in the first phase of this study was representative of VBCPS fourth-grade students. Consequently, the lack of validity coefficients at .30 or above for the first phase can be in part attributed to a lack of responses scored above Level 1.

Unlike fourth-grade students, most seventh-grade students have been able to comprehend that Prompt 2 is asking them to "think outside the box" and determine if additional information may be needed to help them make a choice. Relatively few seventh graders have failed to respond to this prompt. Just over half of the students who took the grade 7 IPT during spring 2014 attained scores of Level 2 or higher in CT2. Based on the results of the second phase of this study, CT2 may be a better indicator than CT1 of critical thinking as operationalized by the CTTST for seventh-grade students. The reverse is true for fourth-grade students, probably because their cognitive abilities are three years behind their older peers. Obtaining a score of Level 2 or above in CT2 requires a more thoughtful, sophisticated response than is required for the same score in CT1. In general, seventh-grade students are better writers than fourth-grade students and thus better equipped to answer all of the IPT prompts. This could be another reason that the correlation value for WC and CT2 was much higher for seventh graders than it was for fourth graders, and also why all of the validity coefficients involving WC and the CCTST scores were greater for the older sample.

SUMMARY

The results of this study provided evidence that the grade 4 IPT and the grade 7 IPT yield scores that are valid for measuring critical thinking. In other words, a student's scores on the IPT can be used to make valid inferences of the student's critical-thinking abilities. However, it is vital that students, parents, educators, policy makers, and the general public recognize that the results of only one assessment should never be used to make high-stakes decisions about individual students. A test simply provides one observation—that of a student's performance on the day that the test was administered—and the score should be used as one component among many to evaluate the student's abilities.

As psychometricians have asserted for years, ''validation is a continual process, one in which an end point is rarely achieved.''[15] Like other assessments, further research with the IPT is

[15]Benson, J. and Clark, F. (1982). A Guide for Instrument Development and Validation. *American Journal of Occupational Therapy*, 36.

warranted to obtain additional information about what the assessment is actually measuring. The results of other versions of the IPT could be correlated with other tests that measure critical thinking, problem solving, or written communication, as well as with other instruments that measure specific domains and content areas. It is recommended that larger samples be employed in future studies with the IPT. By including a sufficient number of students of both genders representing various racial and ethnic groups, it can then be determined if any significant test bias exists with the IPT.

In summary, the present study has offered some evidence of the external validity for the IPT at grades 4 and 7. This outcome lends credibility to the use of the IPT for measuring critical thinking, one of the *Compass to 2015* outcomes for student success. At this time, there are no other VBCPS tests administered divisionwide that assess critical-thinking skills for elementary and middle school students. Accordingly, as Cronbach noted in one of his seminal works, *Test Validation*, "A test of modest validity that provides information not otherwise available is worth using."[16]

Acknowledgements – The author would like to thank the students who participated in both phases of this study as well as the teachers, administrators, and other VBCPS staff who generously offered their assistance. The Office of Research and Evaluation deserves special recognition for assisting with the statistical analyses and offering invaluable advice on earlier drafts of this brief.

[16]Cronbach, L. J. (1971). Test Validation. In R. L. Thorndike (Ed.), *Educational Measurement* (2nd ed.). Washington, DC: American Council on Education.

APPENDIX

Correlations Between IPT Elements

Grade 4 Correlations (n = 207)

IPT Elements	IPT Elements			
	Critical Thinking 1 (n = 206)	Critical Thinking 2 (n = 203)	Problem Solving (n = 204)	Written Communication (n = 204)
Critical Thinking 1	-			
Critical Thinking 2	**.30***	-		
Problem Solving	**.33***	.27*	-	
Written Communication	**.37***	.29*	**.59***	-

Note. Values of .30 or greater are in bold and indicate moderate correlations.
*Correlation is significant at the .01 level of probability (2-tailed).

Grade 7 Correlations (n = 392)

IPT Elements	IPT Elements			
	Critical Thinking 1 (n = 390)	Critical Thinking 2 (n = 391)	Problem Solving (n = 391)	Written Communication (n = 391)
Critical Thinking 1	-			
Critical Thinking 2	**.35***	-		
Problem Solving	.23*	**.33***	-	
Written Communication	**.31***	**.44***	**.56***	-

Note. Values of .30 or greater are in bold and indicate moderate correlations.
*Correlation is significant at the .01 level of probability (2-tailed).