Building a validity framework based on the Standards for Educational and Psychological Testing

Enis Doğan

Research Fellow/Resident Expert at NCES

May 5, 2016

Validity and validation

- "Validity refers to the degree to which evidence and theory support the interpretation of test scores for proposed uses of tests"(p.11).
- Validity is the most fundamental consideration in assessment design, development and implementation.

- Validation: Process of gathering/building validity evidence.
- Validation is an ongoing process (Messick, 1995, p. 740) that is initiated at the beginning of assessment design and continues throughout development and implementation

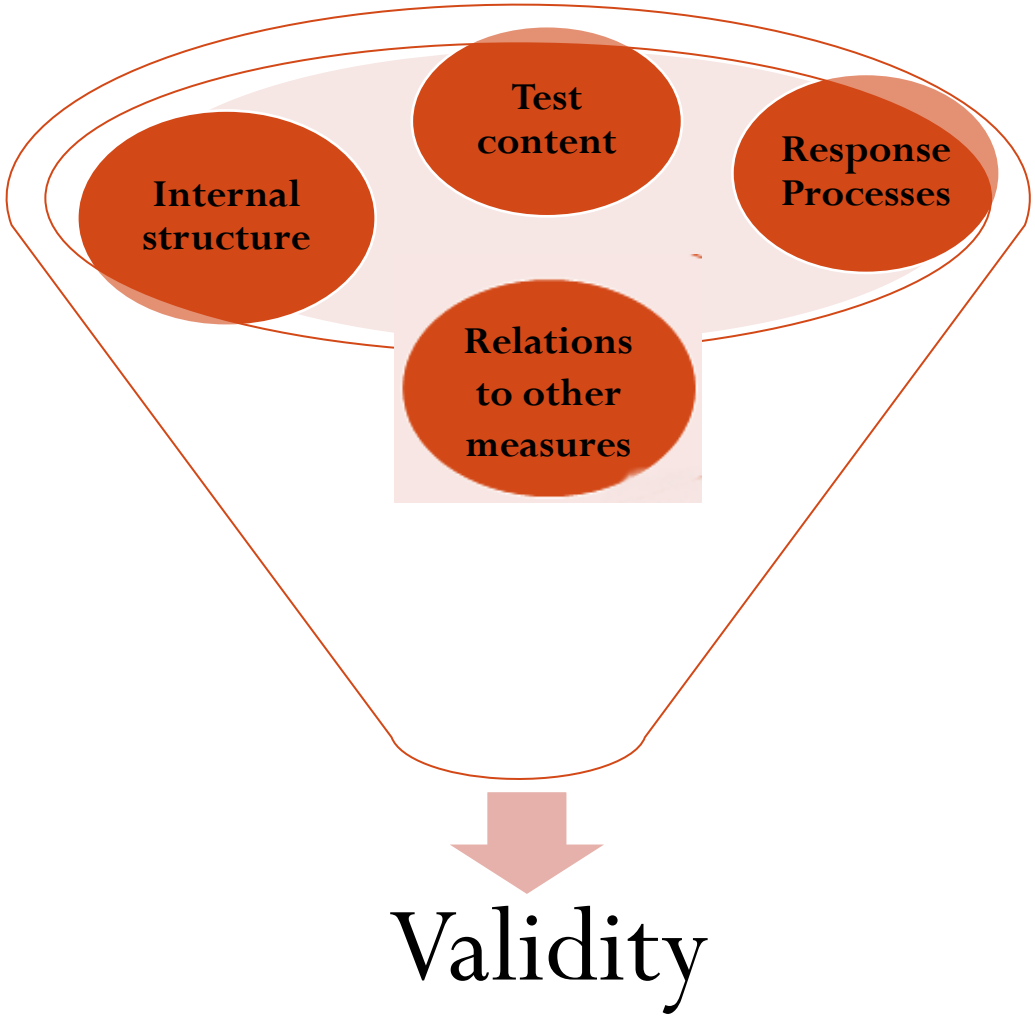
A framework for validity evidence

- **How do we collect validity evidence in a systematic way and avoid gaps for some aspects, or oversaturation for others in terms of validity evidence?**
- As Ferrara (2007) argued, without a framework that can guide the development of validity evidence, it is likely that the full range of validity questions and threats to validity will not be identified.
- Such a framework should help "… expose threats to validity and propose ways to reduce or eliminate these threats" (Haladyna, 2006, p. 739).

A framework for validity evidence

- This illustration is based on Dogan, Hauger, and Maliszewski (2014), where this process was implemented for the PARCC assessments
  1. Identified phases of assessment development and implementation
  2. Listed desired outcomes and conditions that need to be satisfied at each phase (mostly) based on the Standards
  3. Documented empirical and procedural evidence planned to be (or already) collected at each phase while indicating which evidence supports which outcome(s)
  4. Documented source of validity evidence (based on the Standards) for empirical studies

Sources of validity evidence



Phases of assessment development and implementation

- Phase I: Defining measurement targets, item and test development
- Phase II: Test delivery and administration
- Phase III: Scoring, scaling, standard setting
- Phase IV: Reporting, interpretation and use of results

Desired conditions and outcomes: Measurement targets and item development (Phase I)

- **1-A:** The purpose of the assessments is clear to all stake holders.

  Relevant standards: 1.1

- **1-B:** Test specifications and design documents are clear about what knowledge and skills are to be assessed, the scope of the domain, the definition of competence, and the claims the assessments will be used to support.

  Relevant standards: 1.2, 3.1, 3.3

- **1-C:** Items are free of bias and accessible.

  Relevant standards: 7.4, 7.7, 9.1, 9.2, 10.1

- **1-D:** Items measure the intended constructs and elicit behavior that can be used as evidence in supporting the intended claims.

  Relevant standards: 1.1, 1.8, 13.3

Desired conditions and outcomes: Field testing, item banking, and form construction (Phase I)

- 1-E: The item pool as a whole and each test form represents the blueprint and covers the entire range of student performance (including low and high-achieving students).

  Relevant standards: 1.6, 3.2, 3.11, 13.3

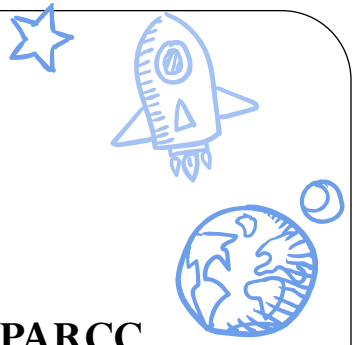
- 1-F: Content of the assessment is rigorous and matches the depth and breadth of CCSS and aligns with the Performance Level Descriptors (PLDs).

  Relevant standards: 3.5

- 1-G: Items with high psychometric quality (e.g., high discrimination/low guessing parameters; high precision; lack of differential functioning) are identified during field testing using representative samples of examinees.

  Relevant standards: 3.3, 3.9, 7.3

Procedural Evidence of Validity: Phase I



- **PARCC's Application for the Race to the Top Assessment Grant (PARCC, 2010)**

  Supported conditions/outcome: 1-A (description of purposes)

- **PARCC Model Content Frameworks (PARCC, 2012)**

  Supported conditions/outcome: 1-B (scope of domain)

- **Performance-Level Descriptors (PLDs) (PARCC, 2013b)**

  Supported conditions/outcome: 1-B (scope of domain)

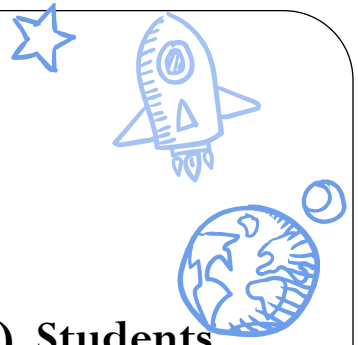
- **Assessment Specifications (PARCC, 2011)**

  Supported conditions/outcome: 1-B (scope of domain), 1-E (blueprint and scale coverage)

- **Cognitive Complexity Framework (Ferrera, et. al., 2014)**

  Supported conditions/outcome: 1-B (scope of domain), 1-E (blueprint and scale coverage)

Empirical Evidence of Validity: Phase I



- **Study 1: Accessibility Studies - English Language Learners (ELLs), Students with Disabilities, and Grade 3 Students (Laitusis, et. al., 2013)**

  Supported conditions/outcome: 1-C (fairness and accessibility)

  Source of validity evidence: Test content, Response processes

- **Study 2: Student Task Interaction Study (Tong & Kotloff, 2013)**

  Supported conditions/outcome: 1-D (intended constructs)

  Source of validity evidence: Response processes

- **Study 3: Quality of Reasoning and Modeling Items in Mathematics (Kotloff, King, & Cline, 2013)**

  Supported conditions/outcome: 1-D (intended constructs)

  Source of validity evidence: Test content, Response processes

- **Study 4: Use of Evidence-Based Selected Response Items in Measuring Reading Comprehension (Pearson, 2013a)**

  Supported conditions/outcome: 1-D (intended constructs)

  Source of validity evidence: Response processes

Desired conditions and outcomes: Assessment Delivery and Administration (Phase II)

- **2-A:** The delivery mode assigned to students does not put any student group (e.g., demographic background, SWD and EL status) at a disadvantage.

  Relevant standards: 13.18

- **2-B:** The directions for test administrators and the test instructions for the students are clear and easy to follow.

  Relevant standards: 3.19, 3.20

- **2-C:** The physical conditions of the testing environment are appropriate for testing

  Relevant standards: 5.4

- **2-D:** Security of test materials and student responses are maintained at all times

  Relevant standards: 5.6. 5.7

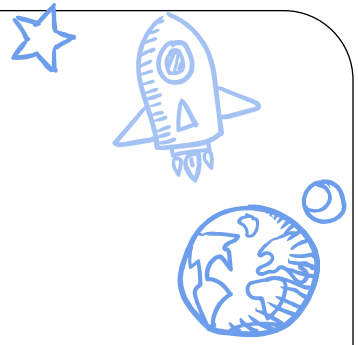
- **2-E:** All students are given the tools (including proper accommodations) they need to indicate their responses accurately and to show what they know and can do.

  Relevant standards: 2.18, 3.15, 5.3, 7.12

- **2-F:** Students are given sufficient time to respond to items and tasks.

  Relevant standards: 2.8, 3.18

Procedural Evidence of Validity: Phase II



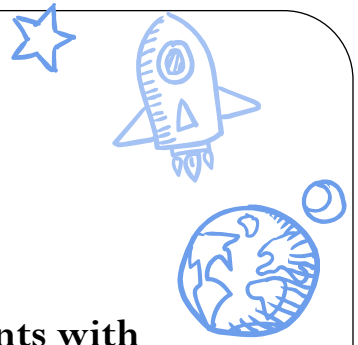
- **Test Administration Manuals** (PARCC, 2014)

Supported conditions/outcome: 2-B (test directions), 2-F (testing time)

- **Accessibility Features and Accommodations Manuals** (PARCC, 2013a)

Supported conditions/outcome: 2-E (response requirements), 2-F (testing time)

Empirical Evidence of Validity: Phase II



- **Study 1: Accessibility Studies - English Language Learners (ELLs), Students with Disabilities, and Grade 3 Students (Laitusis, et. al., 2013)**

  Supported conditions/outcome: 2-E (response requirements)

  Source of validity evidence: Response Processes

- **Study 8: Mode Comparability Study**

  Supported conditions/outcome: 2-A (delivery mode)

  Source of validity evidence: Response Processes

- **Study 9: Device Comparability Study (Strain-Seymour & Davis, 2013)**

  Supported conditions/outcome: 2-A (delivery mode)

  Source of validity evidence: Response Processes

- **Study 10: Quality of Test Administration Instructions Study**

  Supported conditions/outcome: 2-B (test directions), 2-C (testing environment), 2-D (test security), 2-F (testing time)

  Source of validity evidence: Test content

- **Study 11: Text-to-Speech Accommodation Study**

  Supported conditions/outcome: 2-E (response requirements)

  Source of validity evidence: Response Process, Internal Structure

Desired conditions and outcomes: Scoring and Scaling (Phase III)

- **3-A:** Scoring is done reliably and accurately for all types of items and tasks for all summative assessments for all students according to clear scoring rules and rubrics.

  Relevant standards: 1.7, 2.10. 2.13, 3.6, 3.14, 3.22, 3.23, 5.8, 5.9

- **3-B:** Overall scale scores accurately reflect performance on the entire domain through the Performance Based and End of Year assessments.

  Relevant standards: 1.11, 1.12, 2.7

- **3-C:** Scale scores mean the same thing across student groups, forms within a year, and across years.

  Relevant standards: 2.16, 3.6, 4.10, 4.11, 4.17, 5.12, 9.2

- **3-D:** Measurement precision in scale scores is sufficiently high across the scale to support reliable inferences.

  Relevant standards: 2.2, 2.4, 2.14

- **3-E:** Higher scores correspond to higher likelihood of postsecondary success in future.

  Relevant standards: 4.1

- **3-F:** Scale scores allow growth interpretations across years.

  Relevant standards: 4.1

Desired conditions and outcomes: Standard setting (Phase III)

- **3-G:** Cut scores align with the skills and knowledge indicated in PLDs for each level.

  Relevant standards: 1.15, 4.19
- **3-H**: Cut scores are rigorous compared to other relevant national and international benchmarks.

  Relevant standards: 1.15
- **3-I:** Cut scores are vertically aligned across grades.

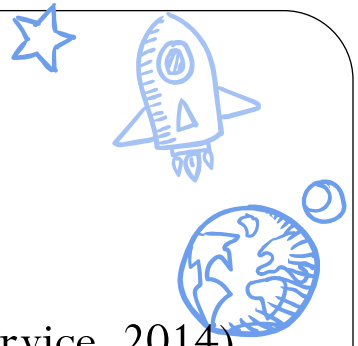
  Relevant standards: 1.15
- **3-J:** The College- and Career-Ready Determination performance level predicts success in postsecondary life.

  Relevant standards: 1.15, 4.19, 4.20, 13.9
- **3-K:** The standard setting panels are representative of all stakeholders.

  Relevant standards: 1.7, 4.21,
- **3-L:** The standard error of measurement is estimated for each cut score

  Relevant standards: 2.15

Procedural Evidence of Validity: Phase III



- **Field Test Psychometric Analysis Plan** (Educational Testing Service, 2014)

  Supported conditions/outcome: 3-C (comparability of scores), 3-D (measurement precision)

- **Technical Memorandum – Standard Setting** (Way, McClarty, & Tong, 2013a)

  Supported conditions/outcome: 3-G (cut scores and PLDs), 3-H (rigor of cut scores), 3-K (standard setting panels), 3-L (standard error of cuts scores), 3-E (predictiveness of scores), 3-I (vertical alignment of cut scores), 3-J (predictiveness of CCR level)

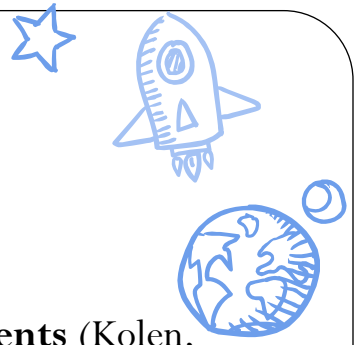
- **White paper – Evidence and Design Implications Required to Support Comparability Claims** (Luecht & Camara, 2011)

  Supported conditions/outcome: 3-C (comparability of scores), 3-D (measurement precision)

- **White paper – Combining Multiple Indicators** (Wise, 2011)

  Supported conditions/outcome: 3-B (scale scores and domain)

Procedural Evidence of Validity: Phase III



- **White paper – Issues Associated with Vertical Scales for PARCC Assessments** (Kolen, 2011)

  Supported conditions/outcome: 3-F (growth interpretations)

- **White paper – Defining and Measuring College and Career Readiness and Informing the Development of Performance Level Descriptors (PLDs)** (Camara & Quenemoen, 2012)

  Supported conditions/outcome: 3-G (cut scores and PLDs), 3-J (predictiveness of CCR level)

- **White paper – Scaling PARCC Assessments: Some Considerations and a Synthetic Data Example** (Brennan, 2012)

  Supported conditions/outcome: 3-D (measurement precision)

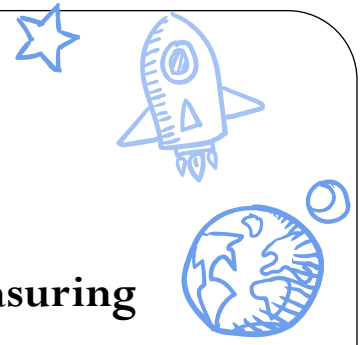
- **White paper – Scores and Scales: Considerations for PARCC Assessments** (Kolen, 2012)

  Supported conditions/outcome: 3-D (measurement precision), 3-F (growth interpretations)

- **Technical Memorandum - PARCC Studies to Examine Comparability of Scores Across States, Assessment Forms, Scoring Methods and Other Relevant Variables** (Thacker, Dickinson, Wise, & Becker, 2014)

  Supported conditions/outcome: 3-C (comparability of scores)

Empirical Evidence of Validity: Phase III



- **Study 4: Use of Evidence-Based Selected Response Items in Measuring Reading Comprehension** (Pearson, 2013a)

  Supported conditions/outcome: 3-A (scoring reliability)

  Source of validity evidence: Response processes

- **Study 5: Use of Narrative Writing Prose Constructed Response (PCR) Tasks in Assessing Reading Comprehension and Writing** (Pearson, 2013b)

  Supported conditions/outcome: 3-A (scoring reliability)

  Source of validity evidence: Response processes

- **Study 8: Mode Comparability Study**

  Supported conditions/outcome: 3-C (comparability of scores)

  Source of validity evidence: Response Processes

- **Study 9: Device Comparability Study**

  Supported conditions/outcome: 3-C (comparability of scores)

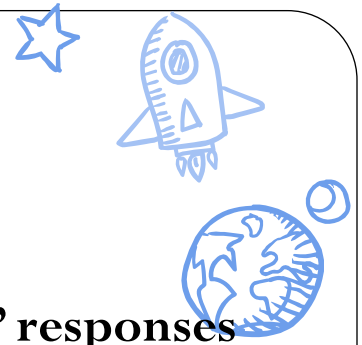
  Source of validity evidence: Response Processes

- **Study 12: Analyses of Field Test Observations and Psychometric Data for Accessibility**

  Supported conditions/outcome: 3-A (scoring reliability)

  Source of validity evidence: Test content

Empirical Evidence of Validity: Phase III



- **Study 13: Validity and Accuracy of Scoring of EL students' responses to PCR Items**

  Supported conditions/outcome: 3-A (scoring reliability)

  Source of validity evidence: Test content

- **Study 14: Automated Scoring Study**

  Supported conditions/outcome: 3-A (scoring reliability)

  Source of validity evidence: Internal structure

- **Study 15: Study of Rubric Choices for ELA/L**

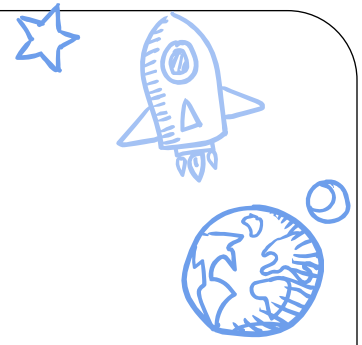
  Supported conditions/outcome: 3-A (scoring reliability)

  Source of validity evidence: Internal structure

- **Study 16: High School Math Comparability Study**

  Supported conditions/outcome: 3-C (comparability of scores)

Empirical Evidence of Validity: Phase III



- **Study 17: Comparability of Assessment Results Study**

  Supported conditions/outcome: 3-A (scoring reliability), 3-C (comparability of scores),

  Source of validity evidence: Internal structure

- **Study 18:Test Administration Mode and Device Study**

  Supported conditions/outcome: 3-C (comparability of scores)

  Source of validity evidence: Internal structure

- **Study 19: International Benchmarking Study**

  Supported conditions/outcome: 3-J (rigor of cut scores)

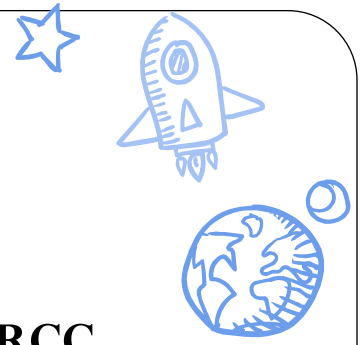
  Source of validity evidence: Internal structure

- **Study 20: Benchmark Study to Inform PARCC Middle and High School Performance Standards**

  Supported conditions/outcome: 3-H (rigor of cut scores), 3-J (predictiveness of CCR level)

  Source of validity evidence: Relations to Other Variables

Empirical Evidence of Validity: Phase III



- **Study 21: Performance of Post-Secondary Students on PARCC Assessments**

  Supported conditions/outcome: 3-H (rigor of cut scores), 3-I (vertical alignment of cut scores), 3-J (predictiveness of CCR level)

- **Study 22: Postsecondary Educators' Judgment Study to Inform Cut Scores in PARCC High Schools Assessments**

  Supported conditions/outcome: 3-J (predictiveness of CCR level), 3-K (standard setting panels)

  Source of validity evidence: Relations to Other Variables

- **Study 23: Longitudinal study of external validity of PARCC performance standards**

  Supported conditions/outcome: 3-H (rigor of cut scores), 3-J (predictiveness of CCR level)

  Source of validity evidence: Relations to Other Variables

Desired conditions and outcomes: Reporting, Interpretation and Use of Results (Phase IV)

- **4-A:** Score reports are developed at student, school, district, state and consortium level featuring relevant comparisons on key indicators such as scale scores, performance level classification, student growth along with standard error for each indicator.

  Relevant standards: 1.11, 1.12, 2.3, 2.7, 4.2, 13.14

- **4-B:** Score reports are accurate and include guidelines in reading and interpreting results and provide actionable results.

  Relevant standards: 2.3, 4.2, 5.10, 5.14, 8.5, 8.6, 5.10, 5.13, 5.16, 7.8,

Evidence of Validity: Phase IV

Procedural Evidence

- **Reporting specifications (PARCC, 2013c)**

  Supported conditions/outcome: 4-A (levels of reporting)

Empirical Evidence

- **Study 24: Prototype Score Report Design Study**

  Supported conditions/outcome: 4-B (accuracy and relevance of score reports)

  Source of validity evidence: Consequences of testing

Conclusions

- There may be other ways to organize the framework
- Some conditions/outcomes will be program specific and some will be applicable to all similar programs
- Best to use this approach in planning (and not just documenting)
- As Ferrara (2007) argued, without a framework that can guide the development of validity research agendas, it is likely that the full range of validity questions and threats to validity will not be identified.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Brennan, R. (2012). Scaling PARCC Assessments: Some Considerations and a Synthetic Data Example. Retrieved from: http://www.parcconline.org/sites/parcc/files/BrennanPARCCScalesWhitePaper.pdf

Buzick, H. (2013). *Accessibility and Fairness Technical Memorandum.* Retrieved from: http://www.parcconline.org/sites/parcc/files/AccessibilityandFairnessTechnicalMemo10-2-13.pdf

Camara, W. & Quenemoen, R. (2012). Defining and Measuring College and Career Readiness and Informing the Development of Performance Level Descriptors (PLDs). Retrieved from: http://www.parcconline.org/sites/parcc/files/PARCCCCRpaperv24-1-8-12CamaraandQuenemoen.pdf

Educational Testing Service. (2014). *Psychometric Analyses of the Field Test Data*. Princeton, NJ: Educational Education Service.

Ferrara (2007). Our Field Needs a Framework to Guide Development of Validity Research Agendas and Identification of Validity Research Questions and Threats to Validity. *Measurement: Interdisciplinary Research and Perspectives, 5(3),* 156-164.

Ferrara, S., Dogan, E., Glazer, N., Haberstroh, J., Hain, B., Huff, K., Larkin, J., Nichols, P.D., Piper, C. (2014). The PARCC Item and Task Cognitive Complexity Code Frameworks: Development, Application, and Validation Evidence Paper to be presented at the annual meeting of the AERA. Philadelphia, PA.

Hain, B. & Piper, C. (2014). PARCC as a Case Study in Understanding the Design of Large-Scale Assessment in the Era of CCSS. Paper presented at the Annual Maryland Assessment Conference. College Park, MD.

Haladyna, T.M. (2006). Roles and importance of validity studies. In Downing, S.M., & Haladyna, T.M. (Eds.), *Handbook of Test Development* (pp. 739-755). Mahwah, NJ: LEA

Kolen, M. (2011). *Issues Associated with Vertical Scales for PARCC Assessments.* Retrieved from: http://www.parcconline.org/sites/parcc/files/PARCCVertScal289-12-201129.pdf

Kolen, M. (2012). *Scores and Scales: Considerations for PARCC Assessments.* Retrieved from: http://www.parcconline.org/sites/parcc/files/KolenPARCCScoresandScales.pdf

Kotloff, L., King, T. & Cline, F. (2013). *Partnership for Assessment of Readiness for College and Careers Cognitive Labs of Mathematics Type II and Type III Items*. Princeton, NJ: Educational Testing Service.

Laitusis, C., Guzman-Orth, D., King, T., Courtney, R. & Cline, F. (2013). *PARCC Item Development Research: Cognitive Labs.* Princeton, NJ: Educational Testing Service.

References

Luecht, R. & Camara, W. (2011). *Evidence and Design Implications Required to Support Comparability Claims*. Retrieved from: http://www.parcconline.org/sites/parcc/files/PARCCWhitePaperRLuechtWCamara.pdf

Messick, S. (1995). Validity of psychological assessment: validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50(9),* 741-749.

PARCC. (2010). *Race to the Top Assessment Application.* Retrieved from: https://www2.ed.gov/programs/racetothetop-assessment/rtta2010parcc.pdf

PARCC (2011). *Blueprints and Test Specifications.* Retrieved from: http://www.parcconline.org/assessment-blueprints-test-specs

PARCC. (2012). *Model Content Frameworks.* Retrieved from: http://www.parcconline.org/parcc-model-content-frameworks

PARCC. (2013A). *PARCC Accessibility Features and Accommodations Manual.* Retrieved from: http://www.parcconline.org/sites/parcc/files/PARCC%20Accessibility%20Features%20and%20Accommodations%20Manual%20November%202013.pdf

PARCC. (2013c). *Section V.D Reporting of the Operational Assessment RFP*. Retrieved from: http://www.parcconline.org/Procurement

Partnership for Assessment of Readiness for College and Careers. (2014). *Test Coordinator Manual.* Retrieved from: http://www.pearsonaccess.com/cs/Satellite?c=Page&childpagename=PARCC%2FpcPALPLayout_v2&cid=1205795411747&pagename=pcPALPWrapper&resourcecategory=Manuals+and+Documents

Pearson. (2013b). *Prose Constructed Response Task Type Research Study.* Iowa City, IA: Pearson.

Strain-Seymour, E. & Davis, L. (2013). *PARCC Device Comparability, Part I: A Qualitative Analysis of Item Types on Tablets and Computers.* Iowa City, IA: Pearson.

Thacker, A., Dickinson, E. Wise, L., & Becker, S. (2014). *PARCC Studies to Examine Comparability of Scores Across States, Assessment Forms, Scoring Methods and Other Relevant Variables Memorandum.* Alexandria, VA: HumRRO

Tong, Y. & Kotloff, L. (2013). *PARCC Student Task Interaction Study.* Iowa City, IA: Pearson.

Way, D., McClarty, K., & Tong, Y. (2013a). *Standard Setting Memorandum: Part 2.* Retrieved from: http://www.parcconline.org/sites/parcc/files/PARCC_StandardSetting_Memo1.pdf

Way, D., McClarty, K., & Tong, Y. (2013b). *Standard Setting Memorandum: Part 2.* Retrieved from: http://www.parcconline.org/sites/parcc/files/PARCC_StandardSetting_Memo2.pdf

Wise, L. (2011). *Combining Multiple Indicators.* Retrieved from: http://www.parcconline.org/sites/parcc/files/PARCCTACPaper-CombiningMultipleIndicatorsRevised09-06-2011.pdf