

SUBJECT INDEX

A

- A-level examinations, 23, 106, 108–109, 121–122, 298
- Access, to educational opportunities: accountability and, 305–306; English language learners and, 191–192, 200, 205; exclusion of low-performing students and, 42, 191–192; policy recommendations for, 319–320; teacher understanding and, 214, 227–228, 234. *See also* Equity
- Access, to higher education: in Singapore, 120–131; in Sweden, 104
- Accountability context: current state performance assessment efforts in, 53–70, 76–85; general acceptance of, 242; impact of, on decisions about students, 304, 312; impact of, on English language learners, 185–186, 191–192; negative effects of, 40–43, 53–54, 260–261, 312, 321; proposal for new state assessment systems in, 85–90; reliability for, 37–38; role of performance assessment in, 47–51, 56, 57–59; school-level, 48–49, 244; student-level, 49, 303–304; task quality for, 76–78; teacher-level, 49. *See also* High-stakes decisions
- Accountability systems: college and career readiness standards and, 304–305, 308; concepts of accountability and, 305–306; developing new, 304–307; elements of, 306–307; of Ohio, 75; policy recommendations for, 320; role of performance assessment in, 56, 57–59; state, 29, 31. *See also* School-level accountability
- Achieve, 321
- “Acid Rain” performance task, 324–326
- ACT, 288
- Active learning skills, development of, through assessment, 103
- Administration, program: local, with central auditing, 82–85, 87–88; mixed centralized/ decentralized, in Australia, 109–110
- Administration, task: costs of, 255; manageability for large-scale, 81–82. *See also* Test-administration costs, expenditures, and benefits
- Administrative support, 317
- Advanced Placement programs, 273–274
- Advanced Placement (AP) Studio Art portfolios, 137
- African American students, automated *versus* human scoring of, 167
- “Air Pocket” performance task, 114
- Alignment: of assessment system to college and career readiness standards, 307–309; with cognitive processes, 172; of rubrics, 153; to standards, 48–49, 76–77, 140–141, 212–214
- Alliance for Excellent Education., 54
- American College Testing, 40
- American Educational Research Association (AERA), 135, 169, 170
- American Institutes of Research, 265
- American Psychological Association (APA), 135
- Analysis-of-errors method, 172
- Analysis-of-reasons method, 172, 177
- Analytic scoring guides, 36
- Analytic scoring procedures, 155, 160–162
- Anchor items, 152, 182–183
- Annual Yearly Progress (AYP), 242
- Architecture licensure exam, 144, 164

- Artificial intelligence, 138, 165, 272
- Arts performance tasks: “creative” type of, 45; of National Assessment of Educational Progress, 45; “responding” type of, 45; sample, from Boston Arts Academy, 224–225; sample, from Washington State, 348–352
- Asian American students, automated *versus* human scoring of, 167
- Assessment(s): authentic, 19–20, 55, 78, 81; categories of, 242–245; challenges for new, 11; for Common Core State Standards, 2–3, 8–11, 259–261, 288–290; continuum of, 292–299; costs and benefits of, 239–257, 259–276; definitions in, 20–22; new opportunities for, 313–315; political debate about role of, 33; reform of education and, 1–5; spending on, in United States, 239–240, 262–267, 377–382, 384*n.* 8:2; in United States *versus* high-performing nations, 93–95, 260. *See also* Large-scale assessment; Performance assessment
- Assessment consortia: in assessment continuum, 294; for Common Core State Standards, 8, 261, 289; as component of high-quality assessment systems, 275, 288, 294; cost savings through, 269, 270, 275–276; expected costs for, 272, 273; standards not measured by, 279–280. *See also* Partnership for Assessment of Readiness for College and Careers (PARCC); Smarter Balanced Assessment Consortium (SBAC)
- Assessment continuum, 292–299, 385*n.* 10:2
- Assessment Solutions Group (ASG), 262–263, 265–266, 377–382
- Assessment system(s): building, for college and career readiness, 277–310, 311–321; continuum of assessments in, 293–299; economical approaches for, 269–272; examples of, 282–287, 291–293; feasibility approaches for, 272–276; to guide student learning, 299–300; high-quality, realizing the benefits of, 268–276; importance of comprehensive, 287–290; multiple audiences and purposes of, 281–283, 290, 299–304; new accountability systems and, 304–307; policy recommendations for, 318–320; recommendations and action steps for, 307–310, 315–318; relative emphasis of purposes for, 290; steps in developing, 291; supports needed for, 308, 317; value of, for individual student decisions, 303–304; value of, to students, 299–304; vision development for, 268–269. *See also* International assessment systems; State performance assessment systems
- Association of Test Publishers, 79, 84
- Auditing, central, 82–85, 87–88
- Australia: assessment systems of, 97–98, 109–119, 284; Curriculum Corporation of, 97. *See also* Queensland, Australia; Victoria, Australia
- Authentic assessment, 19–20, 55, 78, 81
- Automated scoring: adaptive approach in, 250; advances in, 133–134, 163–169; agreement between human scoring and, 166–167; for bad-faith essay detection, 167–168; cost savings through, 169, 270–272, 275; costs and benefits of, 250; effectiveness of, 163–164; features of, 163; issues in, 143–144; scoring algorithms for, 84, 164–165, 169, 272; uses of, 163; validation of, 165–169. *See also* Computer/machine scoring

B

- Bad-faith essays, automated detection of, 167–168
- Basic Education Reform Act (Washington State), 31–32
- BEAR Assessment System, 146–149, 158
- Behaviorist educational philosophy, 383*n.* 2:10
- Benchmark assessments: costs, expenditures, and benefits of, 265–267, 368–372; defined, 242; elements of, 243–245; time-cost savings with, 249. *See also* Interim assessment
- Benchmarks: for curriculum-embedded performance assessment, 87–88; international, 2, 290; student-work, 212–213
- Benefits: of benchmark assessment, 368–372; categories of, 249; for English language learners, 189–191; of formative assessment, 364–367; framework for measuring, 248–249, 364–376; of high-quality assessments, realizing the, 268–276; identification of, 241; measurement of, ways of, 248–249; *versus* multiple-choice tests, 18–20, 42–43, 79–80; opportunity costs and, 246–247, 249–251; of performance assessment *versus* multiple-choice tests, 185–187, 191–192, 264–265; realizing the, in high-quality assessment,

- 268–276; of summative assessment, 373–376; of systems of assessments, 287–290. *See also* Cost-benefit analysis
- Bias: cultural, 198; equating and, 182; evaluation of fairness and, 175–178; guarding against, 213; in human *versus* automated scoring, 167
- BioKIDS project, 148–149, 150
- Blueprints, 317, 320
- Boston Arts Academy, 224–225
- Brookings Institution, 265, 382
- Brown University, Education Alliance at, 68
- Business Information Processing Assessment, 46–47
- C**
- C-rater, 164
- Calibration process: for accuracy, 79, 80; in current performance assessment initiatives, 222; for scoring algorithm development, 165, 272
- California: assessment spending in, 265, 266–267, 378, 379; history of performance assessment in, 32–33; Silicon Valley Mathematics Assessment Collaborative in, 215–220, 233–234; teacher preparation and assessment in, 47
- California Department of Education, 33
- California Learning Assessment System (CLAS), 32–33
- California Standards Test (CST), 33
- Cambridge International Examinations Group, 99
- Cape Cod Lighthouse Charter School, Massachusetts, 229–230
- Carnegie units, alternatives to, 67
- Center for Collaborative Education (CCE), 66, 224
- Certificates of Mastery, 68, 283
- Change measurement: qualitative, 136; quantitative, 136
- Chelsea High School, Massachusetts, 228
- Chemistry domain: learning progressions in, 146–149, 159; scoring rubric for, 156, 158–159
- China, performance assessment in ancient, 27
- Civics performance task, 340–343
- Classroom-based performance tasks: benefits of, 5; for college and career readiness standards, 280–281; as component of comprehensive assessment systems, 282–284, 294–295; for English language learners, 202–204; sample, 340–352; in Washington State, 32, 340–352. *See also* Curriculum-embedded performance tasks; Performance tasks
- Coaching, in performance assessment, 218–219, 221–222, 235
- Cobblestone*, 204
- Cognitive complexity/demand: of Common Core State Standards assessments, 8; and continuum of openness, 141, 172; linking, to criteria, 166; in model-based assessment approach, 142–143; of multiple-choice tests, 6–7; represented as families of tasks, 156; task classification by, 24; validity and, 172
- Cognitive strategies: assessment of college readiness and, 71, 73–74; and measurement of learning progressions, 159
- Cognitive task analysis, 138, 154
- Cognitive theories: in design of learning progressions, 146; in design of performance assessment, 137–139, 140, 146; in design of scoring, 89, 154, 166; validity and, 172
- Collaboration: cost savings through, 269, 270, 275–276; systems approach and, 230–231, 275–276. *See also* State consortia/collaboration; Teacher discussions about student work; Teacher participation
- Collaboration skills, in performance tasks, 135–136
- College admissions, student profiles and, 300–303
- College and career readiness: accountability and, 304–305, 308; assessment initiatives for, 70–76; building assessment systems for, 277–310; Common Core State Standards for, 2–3, 277, 278, 288–290; concerns about, 1–2, 277–278; Connecticut Academic Performance Test (CAPT) for, 62; defining, 279–281; four keys model of, 279, 280; multiple-choice test shortcomings for, 54; recommendations and action steps for, in assessment system, 307–310. *See also* Assessment consortia; Common Core State Standards

- College and Work Readiness Assessment (CWRA): criterion-sampling approach of, 72; overview of, 70–73; stimulus-response complexity of, 24
- College Board, 152
- College Readiness Performance Assessment System (C-PAS), 70–71, 73–74, 360–361
- Collegiate Learning Assessment (CLA): College and Work Readiness Assessment (CWRA) and, 71–73; overview of, 43–44; sample performance task from, 354
- Common Core State Standards: adoption of, 261; assessment criteria for, 2–3, 287–290; assessment initiatives for, 261; assessment system development for, 277–310; importance of instructional alignment to, 77; limitations of, 279; multistate assessment consortia for, 8, 261, 275, 288, 289; overview of, 259–260; performance assessment initiatives for, 215–225; performance assessments for, 8–11, 88; teacher competence in, 207–208; test specifications in, 51–52; validity of measures of, 88–90. *See also* College and career readiness
- Commonwealth Accountability Testing System, 30
- Communication skills, linguistic competency *versus*, 189
- Communities of practice, 213, 223, 228, 235, 240
- Competencies and skills: changes in demand for, 3–4; contemporary requirements for, 3–5; routine, 3, 4. *See also* Higher-order skills assessment
- Composite scores: criterion-sampling approach and, 72; Rhode Island approach to, 69
- Computational linguistics, 165
- Computer-based simulation tasks: automated scoring of, 164; design of, 134, 143–146; examples of, 144–146; features of, 143; issues of, 143–144; task models for, 143, 144
- Computer-based writing skills, 46
- Computer interface familiarity, 143
- Computer/machine scoring: correlations between human scoring and, 44, 166–168, 317; cost savings with, 169, 270–272, 275; in Hong Kong, 125; performance assessment programs with, 71–73, 84; reliability of, 44, 84. *See also* Automated scoring
- Computer skills, sample performance task in, 107, 108
- Conceptual framework, for performance assessment design, 139–141
- Connecticut: assessment programs of, 56, 57, 60–62, 283; No Child Left Behind impact on, 264; sample science task from, 61, 324–326; secondary-level performance assessments in, 56, 57, 60–62; task administrators in, 81–82
- Connecticut Academic Performance Test (CAPT), 56, 57, 60–62, 324–326
- Connecticut Mastery Test (CMT), 56, 60–62
- Connecticut State Department of Education, 56, 60–61
- Consequences, evaluation of, 178–179
- Consistency. *See* Rater agreement; Scoring consistency
- Consortia. *See* Assessment consortia; State consortia/collaboration
- “Constitutional Issues” CBA, 340–343
- Construct-centered performance assessment: for accountability purposes, 49; design framework for, 139–141; scoring method and, 156; scoring rubric and, 155
- Construct-irrelevant variance: in automated scoring, 166; in computer-based assessment, 143; issues of, 136, 139–140; item response theory (IRT) models and, 181–182; language and, 188, 189; sources of, 169–170, 176
- Construct maps, 147, 158, 159, 160
- Construct underrepresentation: in automated scoring, 166; issues of, 139–140; sources of, 169
- Constructed response(s): clear standards for, 22; continuum of complexity of, 23–24; defined, 21; in on-demand component of state assessment systems, 85–86, 88; in online assessment, 71
- Content knowledge, task classification by, 24
- Content map, 142
- Content representativeness, 171
- Content vector analysis (CVA), 165
- Context familiarity, evaluating and ensuring, 176
- Continuous improvement, 309–310
- Controlled assessments, 107
- Corestandards.org, 2, 216
- Corruption-reducing strategies, 49

- Cost-benefit analysis: assessment categories and, 242–245; conceptual framework for, 239–257, 363–376; definitions related to, 245–247; of high-quality assessments, 259–276; relative to multiple-choice tests, 42–43; in standards-based accountability context, 48–49. *See also* Benefits; Costs
- Cost-effectiveness analysis, 241, 262, 318
- Cost savings: through automated scoring, 169, 270–272, 275; through collaborations and consortia, 269, 270, 275–276; through curriculum-embedded assessments, 82, 83; through economies of scale, 252, 254, 269, 270; through formative and benchmark assessments, 249; through local administration, 83; through matrix sampling, 86; methods of, 49, 269–272, 318; by New England Common Assessment Program, 66; through online delivery, 270; in scoring, 272–275, 318; through teacher participation, 272–275
- Costs: analysis of, 251–255, 262–265; assessment categories and, 242–245; for benchmark assessments, 265–267, 368–372; decline in, 43; definitions related to, 245–247; dimensions of, 249–250; economic, 240; expenditures *versus*, 247; of external task administrators, 82; for formative assessment, 364–367; framework for measuring, 251–255, 363–376; GAO study of, 251–252; of high-quality *versus* multiple-choice assessments, 262–265; identification of, 246–247, 249–255; ingredients approach to, 247; of initial instruction *versus* interventions, 249; lessons learned on, 317–318; of meeting NCLB requirements, 27, 239–240, 262–265; opportunity, 240, 246–247; per pupil, 263, 266–267, 271, 378–382, 384*n.* 8:2; of performance assessment *versus* multiple-choice tests, 42–43, 48, 253, 262–265, 317–318; of scoring, 241, 270–275, 317–318, 366; in standards-based accountability context, 48–49; state studies of, 252–254; for student test preparation, 252, 265–267, 377–382, 384–385*n.* 9:2; for summative assessment, 373–376. *See also* Cost-benefit analysis; Expenditures; Teacher time
- Council for Aid to Education, 43, 71
- Council of Chief State School Officers (CCSSO), 11, 17, 79, 84, 261, 268–269
- CPI calculator, 384*n.* 8:1
- Creativity in Action, 259
- Criteria development, 153–155
- Criterion-sampling approach, 72
- Culinary Arts Cook II Assessment, 46
- Culture of teaching, 228–232
- Curriculum: in Australia, 109, 112–113; in Finland, 101; in high-performing nations, 94–95; of International Baccalaureate (IB) Diploma Program, 126; performance assessment alignment to, 137, 139; in Singapore, 121–122; in Sweden, 104
- Curriculum-embedded performance tasks: in College Readiness Performance Assessment System, 73–74; as component of state assessment systems, 87–88; cost efficiencies of, 82, 83; large-scale performance assessment *versus*, 137; in Ohio Performance Assessment Pilot Project, 74; science, 61; scoring of, 209–210; state system of, elements of, 87–88; teacher learning and, 207–235; validity in, 88–90. *See also* Classroom-based performance tasks; Performance tasks
- Cut scores, 303, 309
- ## D
- Deeper learning: building assessment systems for, 277–310; importance of assessment for, 7–8; investment in assessments of, 259–276. *See also* College and career readiness; Higher-order skill assessment
- Demographic differences, research findings on fairness and, 38–39. *See also* Subgroup differences
- Design, performance assessment, 133, 134–149; approaches to, 139; cognitive theories in, 137–139, 140, 146, 166; of computer-based simulation tasks, 143–146; conceptual framework for, 139–141; construct-centered approach to, 139–141, 155; degree of structure in, 141; goal considerations in, 134–137; lessons learned about, 316; to measure learning progressions, 146–149; of scoring rubrics, 153–169; task models for, 141–143, 172

Development costs, expenditures, and benefits: for benchmark assessments, 368; for formative assessments, 364; for performance assessments, 254; for summative assessments, 373

Diagnostic use, of performance assessments: for English language learners, 189–190, 192, 205; *versus* sanctions, 317

Differential item functioning (DIF) studies, 176–177

Directions, for performance tasks, aspects of, 22

Disabilities, students with: performance assessment suitability for, 196; performance issues of, 192. *See also* Special needs students

“Disaster in the Gulf” performance task, 297

Disparity index (DI), 193

Distractor selection differences, 192, 194

Dropout rates, 54, 191

Due notice testing, 77

Dynamic Indicators of Basic Early Literacy Skills (DIBELS), 365

E

E-rater, 164, 165, 168

Economics performance task, 343, 344–345

Economies of scale, 252, 254, 269, 270

EdTPA, 47

Education Alliance at Brown University, 68

Education Policy and Improvement Center, University of Oregon, 73

Educational reform: of 1980s–1990s, 27–28; for college and career readiness, 54; drivers of, 1–5; international, 2; multiple assessment strategies in, 250–251; performance assessment in context of, 1–5, 27–28, 311–321. *See also* No Child Left Behind Act (NCLB)

Educational Testing Service, 2, 17, 152, 165

Educational triage, 42, 191

“Electric Mysteries Performance Assessment,” 186–187

Elementary and Secondary Education Act, 263, 320

Elementary school students, 312

England: A-level examinations of, 106, 108–109; assessment system of, 97, 106–109, 284; General Certificate of Secondary Education (GCSE) of, 106, 107–108, 284; influence of, on other countries’ systems, 106, 120, 121;

National Vocational Qualification assessment of, 106; Qualifications and Curriculum Authority of, 97; test-based school rankings in past, 100. *See also* United Kingdom

English language arts performance tasks: classification of, 23, 24, 25; from England, 107; for English language learners, 189–190, 201–203; of International Baccalaureate (IB) Diploma Program, 127; of National Assessment of Educational Progress, 45–46; from New England Common Assessment Program, 327–329; from New Jersey HSPA/Special Review Assessment, 330–332; from New York Regents Examinations, 333–337; from Partnership for Assessment of Readiness for College and Careers, 9–10, 295; from Smarter Balanced Assessment Consortium, 10. *See also* Writing performance tasks

English language learners (ELLs): benefits of performance assessments for, 189–191; differential subgroup results for, 167; distractor issues and, 192, 194; engagement of, 188, 190, 204, 205; English language support for, 201–202, 205; examples of performance assessments for, 202–204; fairness for, 175–176; impact of linguistic complexity on performance of, 177–178, 188–189, 192–196; instructional improvement for, 200–202; linguistic complexity and, 170, 175–178, 188–189, 192–196; linguistic modification for, 170, 176, 195, 196–198; performance assessments for, 185–205; performance gap of, analysis of, 192–196; policy recommendations for, 319–320; scoring for, 198–199; standardized achievement test shortcomings for, 185–186, 191–192; validity for, 170, 175–176, 192–198

English Language Proficiency (ELP) tests, 201–202, 204

English language support, 201–202

Envision Schools, 297

Equating, 182–183

Equity: accountability and, 305–306; high-stakes testing and, 42, 191–192; policy recommendations for, 319–320; in Singapore, 120; in Sweden, 104; through teacher learning, 214, 227–228, 234. *See also* Access, to educational opportunities

- Error due to raters, 174
- Error due to tasks, 173–174
- Essential Academic Learning Requirements (Washington State), 31–32
- Ethnic bias, 176, 177
- European Commission, 102, 104
- “Evaluating Amelia Earhart’s Life” performance task, 9–10
- “Everyday Life at the End of the Last Ice Age” performance task, 327–329
- Evidence-based model of school finance, 255–256
- Evidence-centered design (ECD), 139
- Exclusion, of low-performers from high-stakes testing, 42, 191–192
- Exhibitions, juried, 298
- Expectations, setting reasonable, 51
- Expenditures: analysis of, 251–255, 262–265; for benchmark assessment, 265–267, 368–372, 377–382; comparative evaluation of, 262–265; costs *versus*, 247; determination of, 240, 245–246, 249–255; dimensions of, 249–250; district-level data on, 246, 377–382; estimates of direct, 256; for formative assessment, 364–367; framework for identification of, 249–255, 363–376; for high-quality assessment (HQA), 271; for performance assessments, 251–255; on standardized multiple-choice assessments, 239–240, 262–267, 377–382; state studies of, 252–254; student test preparation, 252, 265–267, 377–382, 384–385*n. 9:2*; for summative assessment, 373–376. *See also* Cost-benefit analysis; Costs
- Expert review, 134; of automated scoring, 165–166
- Experts, content-domain or education: scoring criteria development by, 153–154; talk-alouds by, 138, 154
- Explanation task model, 142–143
- “Exploring the Maple Copter” performance task, 24
- External performance task administrators: of International Baccalaureate (IB) Degree Program, 126–127; use of, 81–82
- F**
- Fairness: consequences and, 178–179; evaluation of, 175–178; improving, for English language learners, 196–198; issues of validity and, 134, 169–170; lessons learned about, 316; research findings on, 38–39; teacher understanding and, 214
- Feasibility: ensuring, 81, 272; factors in, 272–276
- Federal policy recommendations, 318–320
- Field testing, 134, 149–153
- Finland: assessment system of, 7, 95, 96, 100–103; voluntary matriculation examinations of, 102–103
- Finnish Matriculation Examination, 96, 102
- Finnish National Board of Education, 101–102
- Flexibility option, in No Child Left Behind, 278
- “Floating Pencil” task, 187, 188
- Foreign language skills assessment: in England, 109; in Finland, 102. *See also* English language arts performance tasks; Writing performance tasks
- Formative assessments: costs, expenditures, and benefits for, 364–367; defined, 242; elements of, 243, 244; time-cost savings with, 249. *See also* Performance assessments
- Formulating Scientific Explanations from Evidence* design pattern, 146–149, 150
- G**
- Gaming, 191–192, 304
- Gender bias, 176, 177
- General rubric, 154–155, 156
- Generalizability: construct-centered approach and, 140; content representativeness and, 171; issues of validity and, 173–175; to larger construct domain, 134, 152–153; task models to improve, 141
- Generalized partial-credit model, 181
- Georgia, 265
- Global economy, skills needed for, 3–4
- Goals, of performance assessments, 134–137
- “Golf Balls in Water” performance task, 9
- Gordon Commission on Future Assessment in Education, 2–3, 289–290
- “Got Relieve IT” task, 221
- Graded-response model, 181
- Grading scales: ensuring informative, 80–81; in Finland, 101–102

- Graduate Management Test (GMAT), 168
- Graduate Record Examination, 168
- Graduation decisions and requirements: alternative pathways to, 62–63, 64; assessment system's value for, 303–304; in Australia, 110, 112; in Connecticut, 57; in England, 106–108; in Finland, 102–103; in high-performing nations, 95, 100; in Maine, 59, 67–68; in New Hampshire, 59, 67–68; in New Jersey, 57, 62; in New York, 58, 63–64; in Ohio, 76; in Rhode Island, 59, 68; in Singapore, 122; in Vermont, 70; in Washington, 59
- Graduation portfolios, 283–284
- Graduation projects, 76, 225, 298
- Group differences, research findings on fairness and, 38–39. *See also* Subgroup differences
- H**
- Han dynasty, 27
- Hawaii, explanation task from, 142–143
- Health and fitness performance task, 343, 346–348
- Helium gas balloons simulation task, 145
- Hierarchical rating model, 182
- High-performing nations: assessment features of, 94–100, 260, 284–287; benefits of performance assessment in, 5; educational focus in, 94; on PISA rankings, 94, 100; school-based assessment in, 95, 96, 97, 99; teacher participation in, 208–210, 273–274, 286–287, 314
- High school assessments, state use of performance assessments for, 56–70. *See also* State performance assessment systems
- High School Proficiency Exam (HSPE), Washington State, 32
- High-stakes decisions: in high-performing nations, 95, 100; instructional alignment to standards for, 77; performance assessment for, 48–51; about students, 303–304. *See also* Accountability context; Graduation decisions
- High-stakes testing: exclusion of low-performing students from, 42, 191–192; impact of, on instruction, 40–43, 53–54, 260–261; instructional alignment to standards for, 76–77; with performance assessment, 49–51, 136–137; preparation time and costs for, 252, 265–267, 377–382, 384–385*n.* 9:2; problems with, 27, 41–42, 49–51, 95, 100, 260–261, 278, 312–313; role of performance assessment in, 47–51; in science, 61; shortcomings of, for English language learners, 185–186, 191–192; spending on, 239–240, 262–267, 377–382, 384*n.* 8:2
- Higher-order skill assessment: approaches to, 71; building assessment systems for, 277–310; for college and career readiness, 54, 70–76, 259–261, 277–310; Common Core State Standards-based, 8–11; emerging initiatives for, 70–76; as goal of performance assessment, 135; in high-performing nations, 93–129, 208–210, 260; investment in, 259–276; No Child Left Behind impact on, 6–7; scoring rubrics for, 153; in Singapore, 122; validity in, 88–90
- Hispanic students, automated *versus* human scoring of, 167
- History performance tasks, 203–204, 333–337
- Holistic scoring system, 155, 156, 157, 160–162
- Holt McDougal, 384*n.* 9:2
- Homework, open-ended tasks in, 19–20
- Hong Kong: assessment system of, 124–126, 314–315; Examinations and Assessment Authority, 124–125, 314–315; educational reform in, 124
- Hong Kong Education Bureau, 124, 125–126
- “How Things Work” performance task, 355–359
- Hudson Public Schools, Massachusetts, 228
- Human scoring: accuracy and reliability of, 79–80, 162–163; bias and, 176; cognitive theory applied to, 89; correlations between machine scoring and, 44, 166–168, 317; costs and benefits of, 255; criticism and lack of understanding of, 79; to develop computer algorithm, 84, 165, 169, 272; need for, 35; rater cognition model for, 162–163, 182; rater variability and, 174, 175, 181–183. *See also* Rater agreement; *Scoring headings*; Teacher scorers
- I**
- Illinois Department of Education, 266
- Implementation, of performance assessment programs: lessons learned for, 76–85; local and central administration of, 82–85, 87–88; proposal for new, 85–90; task administration

- issues in, 81–82; task quality issues in, 76–78; technical quality issues in, 35–40, 78–81; technology for, 84–85
- Index number, 242
- Indiana, automated scoring in, 164
- Industrial economy, 3
- Information writing performance task, 327–329
- Innovation Lab Network (ILN), 11, 261
- Institute for Learning, University of Pittsburgh, 210
- Instruction: alignment to, 137; in high-performing nations, 94–95; impact of high-stakes testing on, 40–43, 53–54, 260–261; time spent on interventions in, 249
- Instruction costs, expenditures, and benefits: for benchmark assessment, 370; for formative assessment, 365; for summative assessment, 374
- Instructional improvement: for English language learners, 200–202; evaluation of consequences and, 178–179; in high-quality assessment system, 269; instructional sensitivity and, 179–180; international approaches to, 126; performance assessment impact on, 229–232; performance assessments to support, 202–204, 207–235, 287–288; research findings on, for state performance assessment systems, 40–43, 211
- Instructional sensitivity, 179–180
- Integration of skills, assessment of, 25
- Intellemetric, 164, 168
- Intelligent Essay Assessor, 164
- Intelligent technology, 84–85
- Interdisciplinary skills assessment: construct-irrelevant variance and, 136; in Finland, 102; in Maryland, 135, 136; performance assessment for, 25, 297–298; performance task example for, 297
- Interim assessment costs, 265–267; in California, 266–267, 378, 379; in Kentucky, 266–267, 378, 380; in Massachusetts, 266–267, 381–382. *See also* Benchmark assessments
- Interim measures, 84
- International assessment systems, 5, 93–129; features of, 94–100, 128–129, 284–287; nations studied, 100; overview of, 284–287; reform of, 2; summary table of, 96–99; testing frequency in, 7; US testing *versus*, 6, 7, 93–95, 313
- International Baccalaureate (IB) Diploma Program: assessment system of, 99, 126–128, 284, 298; teacher scoring strategies of, 273–274
- International Baccalaureate Organization (IBO), 126–128
- International competitiveness: concerns about, 1–2; knowledge requirements for, 4–5; multiple-choice tests and, 6; policy discussions about, 94
- International test rankings: high-performing countries in, 94; US performance in, 1, 93–94, 313
- “Interpreting Statistics in the Social Sciences” performance task, 360–361
- Interventions, cost of, 249
- Intraclass correlation, 199
- Item length, 193, 196
- Item response theory (IRT), 73, 75; modeling rater effects with, 181–182; use of, 180–181
- ## J
- Judgment, avoiding, 213
- ## K
- Kappa correlation, 199
- Kentucky: assessment spending in, 253, 266–267, 378, 380; history of performance assessment in, 29–30; impact of assessment reforms in, 41; on-demand testing in, 86; portfolio assessment in, 25, 26, 30, 37, 77–78, 80–81, 83; program administration in, 83; rater consistency in, 37; rich, scorable products of, 77–78; school-based accountability system of, 30; task administrators in, 82; validity findings in, 40
- Kentucky Core Content Test (KCCT), 30
- Kentucky Department of Education, 83
- Kentucky Instructional Results Information System (KIRIS): benefits of, relative to burdens, 265; costs of, 253; overview of, 29–30, 383*n*. 2:4; validity findings of, 40
- Keyboarding, 135
- Knowledge: organization of, expert and novice, 138, 172; rapid expansion of, 4; theories of acquisition of, 137–139. *See also* Content knowledge

L

- Language factors: academic language and, 200; construct-irrelevant, 188, 190; construct-relevant, 188, 198; impact of, on English language learners' outcomes, 177–178, 188–189; performance assessments' effects on reducing, 186–187, 189–191
- Large-scale assessment(s): building, for deeper learning, 277–310; conceptual framework for, 140–141; context of, 55–56; current state performance assessment programs for, 53–70, 76–85, 291–293; design for, 133, 134–149; field testing, 151–153; history of performance assessment for, 26–35; impact of, on instruction, 40–43, 53–54; implementation for, 76–85; performance assessment defined in context of, 20–22; performance assessment in, 47–52, 313–321; policy recommendations for, 318–320; program administration in, 82–85; recommendations for performance assessment in, 51–52, 85–90, 311–321; scoring for, 133–134, 153–169; security issues in testing, 151–153; task administration issues in, 81–82. *See also* International performance assessment; National-level performance assessment; State performance assessment systems
- Latent semantic analysis (LSA), 165
- Learning. *See* Student learning outcomes; Teacher professional development
- Learning certificate, 104
- Learning outcomes. *See* Student learning outcomes
- Learning profiles, 283
- Learning progressions: criteria specification and, 154; defined, 146; design of, 134, 146–149; scoring rubric for, 158, 159; standards and, 209; variation in, 146, 159
- Learning-to-learn skills, 300
- “Letter to the Editor” performance task, 343, 346–348
- Licensure examinations: automated scoring of, 164; computer-based simulations in, 144
- Life skills assessment, 122–124
- Linguistic competency, communicative competence *versus*, 189
- Linguistic complexity/demand: factors in, 196; fairness and, 175–176, 177–178; grammatical, 194–195; impact of, on English language learners, 177–178, 188–189, 192–196; issues of, 192–196; of multiple-choice items, 194–195; performance assessments' reduced, 186–187, 189–191; validity and, 170, 175–176, 177–178
- Linguistic modification: for English language learners, 170, 176, 195, 196–198; example of, 195
- Linking method, IRT, 182–183
- Literary writing, 156
- Little Rock, 249
- Local administration, with central auditing, 82–85, 87–88. *See also* School-based assessment
- Locally sourced foods project, 298

M

- Machine scoring. *See* Automated scoring; Computer/machine scoring
- Maine: in New England Common Assessment Program, 59, 67–68; performance assessment in, 67–68, 283
- Management costs, expenditures, and benefits: for benchmark assessment, 370; for formative assessment, 365; for summative assessment, 375
- Maryland: history of performance assessment in, 30–31, 283; impact of performance assessment in, 41
- Maryland Learning Outcomes, 31
- Maryland School Assessment, 31
- Maryland School Performance Assessment Program (MSPAP): collaborative tasks in, 136; impact of, 41, 179; interdisciplinary tasks of, 135; matrix sampling in, 171, 283; overview of, 30–31; research findings on, 41, 137
- Maryland State Board of Education, 135, 136, 171
- Maryland State Department of Education, 152
- Maryland Writing Test (MWT), 151–152
- Massachusetts: assessment spending in, 266–267, 381–382; task administrators in, 81–82
- Massachusetts Comprehensive Assessment System, matrix sampling in, 86
- “Mastery learning approach,” 67
- Mathematics Assessment Collaborative (MAC), Silicon Valley, 215–220, 233–234

- Mathematics Assessment Resource Service (MARS), 215–220
- Mathematics expertise, qualities of, 25
- Mathematics multiple-choice item, linguistic complexity in, 193–195
- Mathematics performance tasks: classification of, 23, 25; of Connecticut, 61; construct-centered approach to design of, 140; context familiarity of, 176; of England, 108–109; of Finland, 102–103; holistic scoring of, 156, 157; from International Baccalaureate (IB) Diploma Program, 127–128; linguistic demand of, 170, 176, 193–195, 196–198; modification of, for English language learners, 195, 196–197, 198; from Partnership for Assessment of Readiness for College and Careers, 9; from Smarter Balanced Assessment Consortium, 8–9; of Sweden, 105–106
- Mathematics portfolio assessment, 78
- Mathematics scores, validity of, 40
- Matrix sampling: for content representativeness, 171; defined, 86; in Maryland, 171, 283; NCLB and discontinuance of, 26, 86; in New England Common Assessment Program, 86; to offset performance assessment costs, 49; uses of, 86
- Matter strand, in chemistry domain, 146–149, 156, 158–159
- MCAS Pass, 384–385*n.* 9:2
- Meaningfulness, as validity criterion, 173
- Measured Progress, 37, 65, 66, 79
- Measurement-driven instruction, 27–28
- Measurement models, 180–181
- Merit badges, 76, 134
- Minnesota, 283
- Model-based assessment approach, 141–143, 180
- Moderation processes: in Australia, 112, 114, 119; in high-performing nations, 209–210; importance of, 79, 80
- Motivation, student: of English language learners, 188, 190, 204; problem of low, 54
- Multidimensional item response theory (MIRT), 181
- Multiple-choice items: defined, 20, 21; distractors in, 192, 194; English language learners' issues with, 193–195; length of, 193
- Multiple-choice scores: group differences in, 38–39; validity of, 39–40
- Multiple-choice tests: in assessment continuum, 294; cognitive demand deficits of, 6–7; costs of and spending on, 239–240, 262–267, 377–382, 384*n.* 8:2; limitations of, 18–19, 54, 288; performance assessment advantages over, 18–20, 42–43, 79–80; preparation time and costs for, 42, 265–267, 377–382, 384–385*n.* 9:2; problems with high-stakes, 27, 41–42, 53–54, 85, 191–192, 260–261, 312–313; selected response in, 21
- ## N
- Narrowing of the curriculum: concerns with, in 1980s–1990s, 27; due to large-scale high-stakes testing, 41–42, 100, 192, 278
- National Assessment of Educational Progress (NAEP): alignment to standards and performance on, 77; automated scoring of, 164, 317; on English language learners and linguistic complexity, 192; generalized partial-credit model applied to, 181; group differences on scores of, 39, 193–194; mathematics test item of, with linguistic complexity, 193–195; overview of, 33–34; performance tasks of, 44–46; pilot testing of, 17; policy support for, 320; primary trait scoring system of, 155–156; rater consistency in, 37; science simulation tasks of, 145; success of, 50, 179; on teacher support, 52; think-aloud protocol in, 202; validity of, 40
- National Association of State Boards of Education, 54
- National Board for Professional Teaching Standards, 47
- National Center for Education Statistics, 37
- National Center on Education and the Economy, 54
- National Commission on Testing and Public Policy, 17
- National Council of Teachers of Mathematics, 74
- National Council on Education Standards and Testing, 135
- National Council on Measurement in Education (NCME), 135

- National-level performance assessment(s):
for college and career readiness, 70–73; in England, 106–109; examples of current, 43–47, 314; examples of emerging, 70–73, 314; of high-performing nations, 95
- National Occupational Competency Testing Institute (NOCTI), 46–47
- National Research Council, 7, 47, 137, 139, 146, 261
- National Science Foundation, 33
- Natural language processing, 165
- New Basics project, 118–119
- New England Common Assessment Program (NECAP): benefits of, 66; components of, 65; costs of, 253–254; Maine in, 59, 67–68; matrix sampling in, 86; New Hampshire in, 59, 67–68; performance-based components of, 65–66; Rhode Island in, 59, 68–69; sample writing prompt from, 327–329; Vermont in, 28, 69–70, 254
- New Hampshire: assessment framework of, 282, 292–293; graduation portfolios in, 283–284; “mastery of learning” approach of, 67; in New England Common Assessment Program, 59, 67–68; performance assessment in, 59, 67–68, 231–232
- New Hampshire Department of Education, 67, 231, 282
- New Jersey: assessment programs of, 57, 62–63; due notice testing in, 77; High School Proficiency Assessment (HSPA), 57, 62–63, 77, 198, 330; sample performance task from, 330–332; secondary-level performance assessments in, 57, 62–63; Special Review Assessment (SRA), 57, 62–63, 196–197, 199, 330
- New Jersey Department of Education, 62, 63, 199
- New Standards Reference Exam, 28
- New York City, 214, 232
- New York Performance Standards Consortium, 57, 64, 283, 298
- New York Regents Examinations: components of, 58; performance-based components of, 63–64; sample US history performance task of, 333–337
- New York State: assessment programs of, 58, 63–64, 283; secondary-level performance assessments in, 58, 63–64, 283
- New York State Education Department, 63, 64, 79
- Next Generation Science Standards: Ohio Performance Assessment Project mapping to, 220; test specifications in, 51–52
- No Child Left Behind Act (NCLB): costs of, 27, 239–240; English proficiency requirements of, 201; flexibility options for, 278; passage of, 6, 312; public knowledge of, 50, 242; requirements of, 7, 26–27, 211, 242, 263
- No Child Left Behind (NCLB) impacts: on English language learners’ performance, 193; on higher-order skill development and testing, 6–7, 312–313; on matrix sampling, 26, 86; on spending for assessments, 239–240, 262–267, 378, 381–382; on state performance assessment systems, 6, 26–27, 30, 55, 211, 260, 263, 264; on US international comparativeness, 7, 93–94, 313
- North Carolina, 253
- Northwest Educational Assessment, 378
- “Nuclear Power—Friend or Foe?” performance task, 10
- O**
- Objectivity, methods of ensuring, 79
- Occupational program testing, 46–47
- Office of the Superintendent of Public Instruction, 65
- Ohio: accountability system of, 75; performance assessment in, 74–76, 220–223; senior project requirement in, 76
- Ohio Department of Education, 220
- Ohio Performance Assessment Pilot Project (OPAPP), 70–71, 74–76; sample multiday performance task with out-of-class work, 296; sample physics task of, 355–359; sample science task of, 221; teacher participation in, 220–223
- On-demand tests: as component of high-performing nations’ assessment, 95; as component of state assessment systems, 85–86, 88, 282–283
- Online delivery: of College and Work Readiness Assessment, 71–73; of Collegiate Learning Assessment, 44; cost savings through, 270; costs and benefits of, 250
- “Operational Best Practices” (Association of Test Publishers, Council of Chief State School Officers), 79, 84

Oregon: assessment spending of, 265;
Certificates of Mastery of, 283; direct-writing
assessment in, 160–161
Organization for Economic Cooperation and
Development (OECD), 93, 100
Out-of-class performance tasks, 295–298
Outlier responses, scoring of, 84

P

PARCC. *See* Partnership for Assessment of
Readiness for College and Careers
Parents, educating, about performance assess-
ment, 50, 52, 229
Partial-credit model, 181
Partnership for Assessment of Readiness for
College and Careers (PARCC): as compo-
nent of high-quality assessment systems,
275, 289, 294; expected costs of, 272, 273;
overview of, 8, 261; sample performance
tasks of, 9–10, 295; website of (parconline.
org), 10
Paths to proficiency, multiple, 146, 159
Pay-for-performance schemes, 49
Pearson, 378
Peer assessment, 89, 300
Pentucket Regional School District,
Massachusetts, 226, 228–229, 230
Performance assessment(s): advantages of,
versus multiple-choice tests, 18–20, 42–43,
79–80; in assessment continuum, 294–298;
building systems of assessments with, 277–
310, 311–321; classifications of, 22–26; for
Common Core State Standards, 8–11, 215–
225; costs and benefits of, 42–43, 239–257,
259–276; definition of, 12, 20–22; design of,
133, 134–149; elements of, 21–22; for English
language learners, 185–205; evaluation of,
for validity and fairness, 169–180; goals of,
134–137; in high-performing nations, 5,
93–129; for high-stakes decisions, 49–51;
historical overview of, 12, 17–52, 282–283;
increased use of, 314–315; intended score
inferences for, 134, 139; lessons learned
about, 316–317; measurement models and,
180–181; overview of, 11–14; policy recom-
mendations for, 318–320; psychometric
issues in, 134, 180–183; range of activities in,
21; rationale for, 5–8, 27–28; recommenda-
tions for, 311–321; reform context of, 1–5,

27–28, 311–321; research and development
support for, 52, 319; return of, in United
States, 8–11; review of, 149–153; role of,
in standards-based accountability testing,
47–51; scoring of, 133–134, 153–169; at
secondary level, 56–91; status of, 1–11;
task-centered *versus* construct-centered,
49; teacher learning through participation
in, 202–204, 207–235; of teachers, 47. *See*
also International performance assessments;
National-level performance assessments;
State performance assessment systems
Performance Assessment for California
Teachers (PACT), 47
Performance demonstration purpose, 134,
153, 171
Performance measures, defined, 55
Performance Standards Consortium, 64
Performance task(s): in assessment continuum,
294–298; classifications of, 22–26; col-
laborative, 135–136; of Collegiate Learning
Assessment, 43–44, 354; computer-based
simulation, 143–146; construct-centered
approach to, 139–141; as culminating proj-
ect, 298; defined, 20, 21–22; design of, 133,
135–149; development of new generation
of, 52; elements of, 21–22; field testing, 151–
153; for learning progressions, 146–149, 150;
long-term, 296–298; of National Assessment
of Educational Progress, 44–46; of National
Occupational Competency Testing Institute,
46–47; with out-of-class work, 295–298;
quality attributes for, 76–78; state-provided,
87; task-centered approach to, 139, 140; task
demand continuum of, 141; task models for,
141–143, 172
Performance Task Academies, 72–73
Performance task samples: from College
Readiness Performance Assessment
System, 360–361; from Collegiate Learning
Assessment, 354; from England, 108;
from National Assessment of Educational
Progress, 34; from Ohio Performance
Assessment Pilot Project, 355–359; from
Partnership for Assessment of Readiness
for College and Careers, 9–10; sample
new approaches, 353–362; sample state,
323–352; from Smarter Balanced Assessment
Consortium, 8–9. *See also* Arts performance

- Performance task samples (*Continued*)
 tasks; Civics performance task; Economics performance task; English language arts performance tasks; Health and fitness performance task; History performance tasks; Mathematics performance tasks; Science performance tasks
- Persuasive writing, 156, 177
- Physical education standards, 383*n.* 2:1
- Physician computer-based simulation, 164
- Physics performance task, 355–359
- Physics syllabus, 115–118
- Pilot testing, 87
- “Planning a Field Trip” performance task, 8–9
- Policy recommendations, 318–320
- Portfolio assessment: administration of, in large-scale program, 83; in assessment continuum, 298; benefits of, relative to burdens, 264–265; deciding on, 137; definition and elements of, 25–26; graduation, 283–284; in Kentucky, 25, 26, 30, 37, 77–78, 80–81, 83, 265; in mathematics, 78; in New York Performance Standards Consortium, 58, 64; in Queensland, Australia, 118, 286; in Rhode Island, 68, 69; scoring issues in, 26, 36–37, 80–81; in teacher education programs, 25; of teachers, 47; validity of, 40; in Vermont, 28–29, 36–37, 40, 41, 43, 83, 264–265
- Portfolio review, by college admissions, 301
- Practice Planet*, 384*n.* 9:2
- Primary trait scoring system, 155–156
- Principled Assessment Designs for Inquiry (PADI) system, 146–149, 150
- Process skills, task classification by, 24
- Production costs, expenditures, and benefits: for benchmark assessment, 359; for formative assessment, 364; for summative assessment, 373
- Professional development. *See* Teacher professional development
- Professional standards of practice, 305, 306–307
- Profiles: learning, states with, 283; with multidimensional item response theory (MIRT), 181; student, 300–303
- Profiles of Learning, 283
- Program development, recommendations for, 234–235
- Program evaluation costs, expenditures, and benefits: for benchmark assessments, 372; for formative assessments, 367; for summative assessments, 376
- Program in International Student Assessment (PISA), 50, 320; focus of, 94, 103; high-ranking nations in, 94, 100, 120; US ranking in, 93–94, 313
- Project Essay Grader, 164
- Promotion decisions, 303–304
- Protocol analysis, 172
- Pupil-free professional development time, 364–365
- ## Q
- Quaglia Institute, 54
- Qualifications and Curriculum Authority, 104
- Quality, task, 76–78
- Quality Performance Assessment* (Center for Collaborative Education), 224
- Quality Performance Assessment (QPA) network, 66, 224–225, 226–227, 228
- Queensland, Australia: assessment system of, 98, 112–119, 284–286; Core Skills test of, 118; Curriculum, Assessment, and Reporting Framework of, 284–285; New Basics project of, 118–119; physics syllabus of, 113, 115–118; portfolio assessment in, 118, 286; rich tasks approach of, 25, 118–119; science assessments in, 113–114, 119, 285–286; standards and curriculum of, 112–113, 115–118
- Queensland Government, 118
- ## R
- Rand Corporation, 6–7
- Rater agreement: challenges of, 162; factors in, 35–37, 162; between human and automated scoring, 44, 166–168; methods for ensuring, 199; research findings on, 35–37. *See also* Scoring consistency
- Rater cognition model, 162–163, 182
- Rater effects, 181–182
- Rater error/variability, 174, 175, 181–183
- Rater reliability coefficients, 199
- Rater severity/leniency, 181–183
- Raters, errors due to, 174
- Rating scale structure, 181–182. *See also* Scoring headings

- Readiness. *See* College and career readiness
- Reading across the Disciplines tests, 60
- Real-world contexts: authentic assessment and, 19–20, 55; of English performance tasks, 108–109; linkage to, 135; multiple-choice tests *versus*, 19; of Swedish performance tasks, 105–106
- Reliability. *See* Scoring reliability
- Remediation, for consistent scoring, 84
- Reporting costs, expenditures, and benefits: for benchmark assessment, 371; for formative assessments, 366–367; for summative assessment, 375
- Research and development support, 52, 319
- Response dimension, of performance tasks, 23–24
- Response to Literature subtest, 60
- Review process, 149–153
- Rewards and sanctions. *See* High-stakes decisions
- Rhode Island: in New England Common Assessment Program, 59, 68–69; secondary-level performance assessment in, 59, 68–69, 283
- Rhode Island Board of Regents, 68
- Rhode Island Department of Education, 68
- Rich, scorable products, 77–78
- Rich tasks: banks of, 25, 119, 292; defined, 118, 141; extended performance tasks and, 296–299; in Queensland, Australia, 25, 118–119
- Rubrics. *See* Scoring rubrics
- S**
- Saltus model, 159
- San Diego, California, 232
- SBAC. *See* Smarter Balanced Assessment Consortium
- Scaffolding, 149, 202
- School-based assessment: in Australia, 109–110, 112; in England, 106; in Finland, 101; in high-performing nations, 95, 96, 97, 99, 209; in Hong Kong, 124–125; in International Baccalaureate (IB) Diploma Program, 126; in Singapore, 124; in Sweden, 103–104. *See also* Local administration, with central auditing
- School finance, evidence-based model of, 255–256
- School-level accountability: Hong Kong system of, 125–126; in Kentucky, 29, 77–78, 82; in Maryland, 31; matrix sampling for, 171, 283; policy recommendations for, 320; use of performance assessment for, 48. *See also* Accountability systems
- School rankings, high-performing nations' assessments and, 95, 100
- Science expertise, qualities of, 25
- Science knowledge integration scoring rubric, 159–160
- Science performance tasks: classification of, 23, 24, 25; computer-based simulation, 145–146; of Connecticut, 61, 324–326; for high-stakes assessment, 61; learning progressions in, 146–149, 150; linguistic demand of, 186–187; of Maryland, 135; of National Assessment of Educational Progress, 34, 45, 202; of Ohio Performance Assessment Pilot Project, 355–359; of Queensland, Australia, 113–114, 119, 285–286; scoring procedures for, 161; scoring rubrics for, 156, 158–160; of Singapore, 122–123; task sampling variability of, 175; teacher learning from, 222–223; of Victoria, Australia, 110–111
- Score inferences: design and, 134, 139–140; validation of, 169–180. *See also* Generalizability; Validity
- Score interpretations and use: in computer-based assessment, 143–144; deciding on, 134; design and, 134, 139; equating and, 182–183; item response theory (IRT) models for, 180–182; linking and, 182–183; proper, 317; validity and fairness of, 134, 169–180
- Score levels, 154
- Scoring: analytic, 155, 160–162; costs of, 241, 270–272, 317–318, 366; criteria in, 153–155, 212–214; development of, 133–134, 153–169; for English language learners, 198–199; holistic, 155, 156, 157, 160–162; lessons learned about, 316; of portfolio assessments, 26, 36–37, 80–81; primary trait, 155–156; procedures, 155–162; teacher learning from, 84, 199, 226–228. *See also* Automated scoring; Human scoring; Validity
- Scoring consistency: methods to ensure, 84, 162–163; in portfolio assessment, 26, 36–37; primary trait rubrics and, 156; research findings on rater agreement and, 35–37. *See also* Rater agreement

- Scoring costs, expenditures, and benefits: for benchmark assessment, 371; for formative assessment, 366; for performance assessment, 255; for summative assessment, 375
- Scoring inflation, 80–81, 83
- Scoring orientation, 212
- Scoring reliability: factors in, 38; methods for establishing, 199; problems with, 78–79; research findings on, 37–38; of teacher scorers, 79–80, 162–163
- Scoring rubrics: cognitive task analysis for, 138; criteria in, 153–155, 212–214; design of, 79, 153–169; for English language learners, 199; field testing, 79, 151–153; general, 154–155, 156; teacher calibration with, 210, 222
- Scoring sessions: benefits of, 212–215, 223, 226–228; elements of, 212; learning from, 226–227. *See also* Teacher discussions about student work
- Security maintenance, in large-scale field testing, 151–153
- Selected response, defined, 21
- Self-fulfilling prophecies, 213
- Self-reflection, student, 135–136
- Senior projects, 76, 225, 298
- Silicon Valley Mathematics Assessment Collaborative (MAC), 215–220, 233–234
- Simulations. *See* Computer-based simulation tasks
- Singapore: assessment system of, 99, 120–124, 284, 298; educational reform in, 2, 120; Examinations and Assessment Board (SEAB), 99, 121, 124; General Certificate of Examinations (GCE) of, 99, 121; Institute of Education of, 124; Primary School Leaving Examinations of, 121; Project Work (PW) of, 122–124; Strategies for Active and Independent Learning (SAIL) of, 124; Thinking Schools, Learning Nation initiative of, 122
- Situation, structured, 21–22
- Smarter Balanced Assessment Consortium (SBAC): as component of high-quality assessment systems, 275, 289, 294; expected costs of, 272, 273; New Hampshire in, 67, 292; overview of, 8, 261; sample performance tasks from, 8–9, 10; website of (smarterbalanced.org), 10
- “Snack Time” visual arts performance task, 349–352
- Social studies performance tasks, 203–204
- Souhegan High School, New Hampshire, 226–227
- Special needs students, alternative assessment pathways for, 62–63. *See also* Disabilities, students with; English language learners
- Stakeholders: assessment data for different, 281–283, 299–304; educating, about performance assessment, 50, 52; working with, 309
- Standardized achievement tests: in assessment continuum, 294; costs of and spending on, 239–240, 262–267, 377–382, 384*n.* 8:2; preparation time and costs for, 252, 265–267, 377–382, 384–385*n.* 9:2; shortcomings of, 185–186, 191–192, 288; state budgets for, 262; widespread acceptance of, 242, 243, 244. *See also* High-stakes testing; Multiple-choice tests; Summative assessments
- Standardized Testing and Reporting (STAR) exams, 33
- Standards: accountability and, 304–305; aligning performance assessment to, 48–49, 76–77, 137, 140–141; aligning scoring with, 212–214; of high-performing nations, 95, 209; internationally benchmarked, 2, 290; in Queensland, Australia, 112–113, 115–118; reform of, drivers of, 1–5; revising, 51; subject domain differences and, 48. *See also* Common Core State Standards; Next Generation Science Standards
- Standards for Educational and Psychological Testing* (American Educational Research Association), 70, 135, 169, 303
- Stanford Achievement Test (SAT), 33, 152, 288
- Stanford Center for Assessment, Learning and Equity (SCALE), 220
- State consortia/collaboration: cost savings through, 269, 270, 272, 273; for performance task development, 52. *See also* Assessment consortia; New England Common Assessment Program
- State performance assessment systems: building, for deeper learning, 227–310, 311–321; current and emerging, 53–70, 76–85, 215–225, 283–284, 291–293; demise of past, 6, 17–18, 26–27, 34–35, 55, 78–79, 211, 260, 263, 264; high-stakes problems in, 50; historical overview of (1990s-era), 5–6, 17–18, 26–35, 283–283; impact of past, 40–43, 274;

- instructional improvement from, 6, 274;
 lessons learned from current and emerging,
 76–85; NCLB impact on, 6, 26–27, 55, 211;
 number of states with, 26; on-demand com-
 ponent combined with, 85–86, 88; proposal
 for new, 85–90; purposes of, 56; research
 findings on past, 6, 35–43, 211; sample per-
 formance tasks from, 323–352; secondary-
 level, 56–70; states with current, 56–70, 215,
 283–284; student achievement improvement
 from, 6; teacher participation in, 211–225,
 314; technical quality of, 35–40, 78–81
- State policy recommendations, 318–320
- Stimulus materials: continuum of complexity
 of, 23–24; limitations of multiple-choice, 19;
 as performance task element, 21, 22
- Stimulus-response classification scheme, 23–24
- Stipends, 274
- Structure, degrees of, 141
- Student boredom, 54
- Student choice, 136
- Student learning outcomes: accountability
 for, 307; evaluation of consequences and,
 178–179; in high-quality assessment system,
 269; investment in assessments for, 259–276;
 multiple assessment strategies and, 250–251,
 299–300; from teacher participation in
 performance assessment, 219, 229–232,
 287–288
- Student profiles, 181, 283, 300–303; compo-
 nents of, 302
- Student self-assessment, 89, 300
- Student test preparation time and costs, 252,
 265–267, 377–382, 384–385*n. 9:2*
- Students: exclusion of low-performing, from
 high-stakes testing, 42, 191–192; value of
 high-quality assessments to, 299–303. *See*
also English language learners
- Study Island*, 384*n. 9:2*
- Subgroup differences: analysis of, 89, 176–178,
 192–196; due to validity problems, 170,
 175–176, 192; in human *versus* automated
 scoring, 167; research findings on fairness
 and, 38–39. *See also* English language learn-
 ers (ELLs)
- Subject field: content standard analysis by, 48;
 task classification by, 24–25
- Summative assessments: costs, expenditures,
 and benefits of, 373–376; defined, 242–243;
 elements of, 244, 245. *See also* Multiple-
 choice tests; Standardized achievement tests
- Support experts, for New England Common
 Assessment Program states, 67
- Support materials, 52; in New England
 Common Assessment Program, 66,
 69–70; for state system of curriculum-
 embedded performance assessment,
 87, 88. *See also* Teacher-scorer support;
 Vendors
- Sweden: assessment system of, 96, 103–106;
 National School Board examinations of,
 105–106
- Swedish National Agency for Education,
 104
- System learning, 309–310
- ## T
- Task-based language assessments (TBLAs),
 189
- Task-centered design approach, 139, 140
- Task demand continuum, 141
- Task difficulty, 181–182
- Task models, 141–143, 144, 172
- Task number: generalizability and, 173–174;
 score reliability and, 38
- Task quality, 76–78
- Task sampling variability, 173–175
- Task shells. *See* Task models
- Task specificity, 173–174
- Tasks. *See* Performance tasks
- Teacher discussions about student work: ben-
 efits of, 70, 162, 210, 213–215, 226–228; in
 current US performance assessment initia-
 tives, 215–225; scoring sessions for, 212–215,
 226–227; in Vermont, 70
- Teacher participation, 207–235; in Australia,
 110, 112–113; benefits of, 211–215, 226–235,
 272–275, 287–288, 314; in current US
 performance assessment initiatives,
 215–225; in Finland, 101; in high-
 performing nations, 208–210, 286–287;
 in high-quality assessment system, 269,
 272–275; in Hong Kong, 125; outcomes of,
 219, 229–232; in Sweden, 104–105. *See also*
 Teacher scorers
- Teacher preparation programs, 47, 100–101
- Teacher professional communities, 213, 223,
 228, 235, 240

- Teacher professional development and learning: assessment in, 47; in current US performance assessment initiatives, 215–225; for formative assessment, 243; on higher-order skill instruction and assessment, 72–73, 207–208; as incentive for scoring duties, 274, 364–365; lessons learned on, 317; performance assessment participation as support for, 207–235; from performance assessment scoring, 84, 199, 226–228; school-based structures for, 70; short-term models of, 208; in Singapore, 120. *See also* Instructional improvement
- Teacher-scorer support: for consistency and comparability, 84; need for, 52; in New England Common Assessment Program states, 66, 67, 69–70; in Ohio Performance Assessment Project, 221–222; in Silicon Valley Mathematics Assessment Collaborative, 216–217
- Teacher scorers: cognitive theory used by, 89; cost savings with, 270; costs and benefits of, 255, 272–275; in current US performance assessment initiatives, 215–225; in high-performing nations, 209–210, 286–287; in New England Common Assessment Program, 66, 68, 69; in New York State, 64; stipends for, 274; teaching and learning improvement through use of, 207–235; in the United States, 211–215; in Vermont, 69–70
- Teacher time: allocation and integration of, 241, 256–257, 274; components of, 240–241; for formative assessment, 364–365; on interventions *versus* initial instruction, 249; weight of, 240, 256–257
- Teachers: pay-for-performance schemes for, 49; performance assessment of, 47; professional standards of practice for, 305, 306–307
- Teaching. *See* Instruction; Instructional improvement
- Technical quality: of College Readiness Performance Assessment System, 73–74; of current and emerging state performance assessment programs, 78–81; of Ohio Performance Assessment Project, 75; of past state performance assessment programs, 35–40
- Technology: cost savings through, 275; lessons learned about using, 316–317; for performance assessment systems, 84–85, 316–317
- Templates. *See* Task models
- Test-administration costs, expenditures, and benefits: for benchmark assessment, 370; for formative assessment, 365; for performance assessment, 255; for summative assessment, 374. *See also* Administration, task
- Test of English as a Foreign Language, 189
- Test specifications, 140, 144
- Thatcher government, 100
- Think-aloud protocol, 202
- ThinkReady Assessment System (EPIC), 294
- Time costs. *See* Teacher time
- TIMMSS (Trends in International Mathematics and Science Study) assessments, 120
- Title I, 77, 201
- Title III, 201
- Tools for Teachers*, 218
- Traffic flow performance task, 296–297
- Training: in current US performance assessment initiatives, 218–219, 221–222; for performance assessment, 84, 199, 235. *See also* Teacher professional development and learning
- Training costs, expenditures, and benefits: for benchmark assessment, 369; for formative assessment, 364–365; for summative assessment, 374
- Transcripts, 301
- Transdisciplinary learnings, 25. *See also* Interdisciplinary skills assessment
- Transparency, as validity criterion, 173
- Trend scoring, 183
- Triage, educational, 42, 191
- Triangulation of measures, 309
- 21st Century skills, 3–4, 55, 259; for all ages, 312; building assessment systems for, 277–310, 311–321; Common Core State Standards and, 259–260; high-performing nations’ focus on, 94, 260; PISA assessments’ focus on, 94. *See also* College and career readiness; Common Core State Standards; Higher-order skills assessment
- U**
- United Kingdom: examinations in, 23; National Assessment of Educational Progress tasks and, 33; task administration in, 81. *See also* England “United States History and Government Document-Based Essay,” 333–337
- US Bureau of Labor Statistics, 384*n.* 8:1
- US Department of Education, 29, 278, 311

US General Accounting Office (GAO), 240, 263–264; cost analysis study of, 251–252
University of California, Berkeley, 4
University of Oregon, Education Policy and Improvement Center, 73
University of Pittsburgh, Institute of Learning, 210
University of the State of New York State Education Department., 64

V

Validity: of analytic *versus* holistic scoring, 160–162; of automated scoring procedures, 165–169; chain of evidence for, 89; cognitive complexity and, 172; in computer-based assessment, 143–144; consequential evidence for, 178–179; content representativeness and, 171; criteria for, 170–180; defined, 169; design and, 139; for English language learners, 170, 175–176, 192–198, 199; evaluation of, 134, 169–180; factors in establishing, 39–40, 73–74, 88–90, 170–180; generalizability and, 173–175; for higher-order thinking assessment, 88–90; instructional sensitivity and, 179–180; linguistic complexity and, 170, 175–176, 177–178, 193–198; meaningfulness and, 173; new approach to, 88–90; problems with, 78–79; research findings on, 39–40; of standardized achievement tests, 185; theory, 89; threats to, 169–170; transparency and, 173
Values conflicts, 50, 79
Variance components, 174
Vendors: interim test materials of, 378, 384–385*n.* 9:2; for scoring, 274
Vermont: current assessment system of, 69–70; Developmental Reading Assessment (DRA), 28–29; history of performance assessment in, 28–29, 283; in New England Common Assessment Program, 28, 69–70, 254; rater consistency in, 36–37; School Quality Standards mandate of, 69; secondary-level performance assessment in, 69–70; Summer Auditing Institute, 29; validity in, 40
Vermont Department of Education, 70
Vermont Portfolio Assessment Program: administration of, 83; historical overview of, 25, 26, 28–29; impact of, 41, 43, 264–265
Vermont State Board of Education, 69
Victoria, Australia: assessment system of, 98, 110–112; Curriculum and Assessment Authority (VCAA) of, 99, 110, 112

Visual arts classroom-based assessments, 348–352. *See also* Arts performance tasks

W

Washington: assessment consequences in, 178–179; assessment programs of, 59, 64–65; history of performance assessment in, 31–32; Measurements of Student Progress (MSP), 32, 65; secondary-level performance assessments in, 31–32, 59, 64–65
Washington Assessment of Student Learning (WASL): components of, 59, 64–65; overview of, 31–32; sample classroom-based assessments from, 340–352
Washington Post, 31
Washington State Institute for Public Policy, 32
Webb Depth of Knowledge framework, 6
“What you test is what you get” (WYTIWYG), 27–28
Wireless Generation website, 365
Workforce representatives, 309
Writing across the Disciplines assessment, 60–61
Writing performance tasks: automated scoring of, 164–165, 166, 167–168; of Collegiate Learning Assessment, 44, 354; for English language learners, 190–191; of International Baccalaureate (IB) Diploma Program, 128; for language arts, 25; for large-scale high-stakes contexts, 136–137; of National Assessment of Educational Progress, 45–46; from New England Common Assessment Program, 327–329; from New Jersey HSPA/Special Review Assessment, 330–332; from New York Regents Examinations, 333–337; rater cognition model for scoring, 162–163; rater consistency for, 35; score reliability for, 38; scoring procedures for, 155–156, 160–162. *See also* English language arts performance tasks
Wyoming, 249, 283

Y

“You and the Economy” performance task, 343, 344–345

Z

“Zoo Mug” performance task, 348–349

