Value-Added Evaluation in Higher Education: Models and Methods

Authors: Yang Xue, Li Jing, Yang Xue, Li Jing

Date: 2024-10-15T00:00:00+00:00

Abstract

From a methodological perspective, this study systematically reviews the development trajectory of domestic and international value-added evaluation research and related evaluation models, with the aim of providing reference for the reform practice of higher education evaluation and graduate education quality evaluation in China.

Full Text

Preamble

Value-Added Assessment in Higher Education: Models and Methods

YANG Xue, LI Jing
Information Management School, Sun Yat-sen University, Guangzhou, Guangdong 510006

Abstract: From a methodological perspective, this study systematically reviews the development trajectory and relevant evaluation models of value-added assessment research both domestically and internationally, aiming to provide references for the reform of higher education evaluation and graduate education quality assessment in China.

Keywords: Value-added evaluation; Evaluation system; Educational reform; Graduate education quality assessment

As educational reform practices deepen, the limitations of traditional evaluation models—characterized by excessive instrumentalism and managerial attributes that diverge from the essential nature of education and educational evaluation—have become increasingly prominent. The academic and educational communities urgently need to explore evaluation mechanisms aligned with the high-quality education system of the new era. To enhance the scientific rigor, professionalism, and objectivity of educational evaluation, the Central Committee of the Communist Party of China and the State Council issued the "Overall Plan for Deepening the Reform of Education Evaluation in the New Era" in October 2020, explicitly proposing four new initiatives: "improving outcome evaluation, strengthening process evaluation, exploring value-added evaluation, and improving comprehensive evaluation" [1]. This marked the first time that value-added evaluation was formally established in China's educational policy documents. Against this backdrop, academic research on value-added evaluation has surged. While scholars have conducted extensive theoretical discussions on the internal mechanisms and implementation pathways of educational value-added evaluation, systematic research findings on evaluation models and methods remain limited. Therefore, this paper aims to systematically review the development trajectory and relevant evaluation models of value-added assessment research from a methodological perspective, providing references for the reform and practice of higher education evaluation and graduate education quality assessment in China.

Funding: This research is supported by the 2024 Guangdong Provincial Graduate Education Innovation Program Project (Project No.: 2024JGXM_{004}) and the 2024 Graduate Education Research Cultivation Project Series of the China Association of Graduate Schools.

Author Information: YANG Xue (1999—), female, Han ethnicity, from Liaoning, master's student, research interests: big data analysis in science and technology, 18664777121, yangxue2115@163.com, No. 132 Waihuan East Road, University Town, Panyu District, Guangzhou, Guangdong; LI Jing (1985—), female, Han ethnicity, from Shandong, Ph.D., associate professor, doctoral supervisor, research interests: big data analysis and decision-making in science and technology, information users and services, 18802093621, lijing359@mail.sysu.edu.cn.

2.1 Origin and Development of Value-Added Assessment

Value-added assessment originated in the 1960s from research on school performance issues in American education. The 1966 Coleman Report, submitted to the U.S. Congress, catalyzed the emergence of value-added assessment models. The report argued that schools' influence on student academic progress was far smaller than factors such as family background, social context, and student intake conditions, contending that traditional evaluation models focusing solely on outcomes could not fairly reflect the net effect of schools with disadvantaged student populations on academic growth. This conclusion triggered widespread reflection on school effectiveness and evaluation methods [2]. Following this report, research on value-added performance evaluation of schools and teachers developed rapidly in Western countries, particularly the United States. However, due to statistical technique limitations, value-added methods were not fully applied until the mid-1980s. By the late 1980s, optimized statistical techniques provided more precise, reliable, and operational analytical methods for value-added assessment, enabling its rapid implementation and expanding both the depth and breadth of research [3]. The 1983 report "A Nation at Risk: The Imperative for Educational Reform" formally established value-added assessment as part of the education evaluation system and initiated its practical application [4]. With the launch of the Tennessee Value-Added Assessment System and the signing of Tennessee's Education Improvement Act in 1992 [5], value-added assessment models were first fully implemented in Tennessee's educational reform practices, achieving positive results. Influenced by this success, other U.S. states and Western countries actively introduced value-added assessment models into their educational practices. In 2002, the federal education law "No Child Left Behind" explicitly required measuring school or teacher performance based on students' annual progress, thereby providing legal support and ensuring large-scale implementation across the United States [6].

Overall, the theoretical and methodological system of value-added assessment has gradually developed and matured through successive educational reforms in the United States, experiencing stages of emergence, implementation, development, large-scale promotion, and continuous improvement. However, value-added assessment was introduced to China's education sector relatively late, and its development has been comparatively slow. Introduced around the 1990s, Chinese scholars have continuously explored its characteristics, foundational theories, and methods, yet systematic theoretical achievements and mature methodological frameworks have not been established. It was not until 2020, with the release of the "Overall Plan for Deepening the Reform of Education Evaluation in the New Era," which clarified the position of value-added assessment in the education system, that value-added assessment research entered a period of vigorous growth in China.

2.2 Theoretical Connotations and Characteristics of Value-Added Assessment

The concept of "value-added" originates from economics, referring to the incremental portion when comparing inputs to outputs [7]. Applied to education, it denotes the degree of student progress after receiving school education. Value-added assessment typically calculates this increment based on the "profit = output - input" logic by tracking and comparing longitudinal student development data over a specified period. Compared with traditional evaluation models, value-added assessment offers innovative conceptual advantages. Traditional models evaluate quality based on a single assessment result [8], whereas value-added assessment emphasizes a "development" philosophy, determining progress levels by measuring the distance between students' starting and ending points [9]. Thus, value-added assessment is essentially a longitudinal, developmental evaluation model that better facilitates incentive mechanisms, taps student potential, and promotes comprehensive student development and progress. Additionally, it can preliminarily exclude the influence of potential factors beyond the control of evaluation objects [10], thereby ensuring relative fairness.

These characteristics yield several distinct advantages. First, by centering on student development, value-added assessment returns to the essential nature of educational activities and promotes educational equity. Under traditional models, teachers and schools tend to focus more on high-achieving students while neglecting lower-ranked students, since excellent students typically produce high scores that boost institutional rankings. In contrast, value-added assessment measures actual progress based on improvement values rather than baseline conditions, significantly diminishing the ranking advantage conferred by high-achieving students. Consequently, teachers must attend to all students' progress to achieve high evaluation rankings. This ensures that all students receive equal attention from schools and teachers, demonstrating value-added assessment's unparalleled superiority in promoting educational equity and student development. Second, value-added assessment emphasizes multidimensional perspectives on student development, encouraging students to enhance comprehensive competencies alongside academic performance. Existing research indicates that value-added assessment models can also provide unbiased predictions of value-added performance, yielding more objective and precise results [11] that serve as effective tools for monitoring student development quality, evaluating teacher performance, optimizing team structures, and improving teaching quality [12].

3 Value-Added Assessment System

The fundamental value-added assessment system comprises three elements: evaluation objects, evaluation data, and scientific value-added assessment methods. The following sections analyze each component in detail.

3.1 Value-Added Assessment Objects and Data

Initially, value-added assessment focused primarily on schools and teachers. As national and educational attention to students' holistic development has increased, research gradually expanded to explore student value-added and personalized development plans, making schools, teachers, and students the three primary evaluation objects. Data describing student progress serves as the source for evaluating these subjects' value-added contributions, primarily obtained through three instruments: coursework examination scores, standardized tests, and value-added assessment scales. Coursework examination scores and standardized tests constitute direct assessments of student value-added, while value-added assessment scales represent indirect assessments.

Standardized tests focus on measuring students' core cognitive abilities through specifically designed questions. Practically, tests can be administered in stages according to evaluation needs, with student progress quantified by score differences between initial and final assessments. The United States has a long tradition of standardized testing, developing classic instruments such as the Measure of Academic Proficiency and Progress (MAPP), the Collegiate Learning Assessment (CLA), and the Basic Skills Test through educational reform practices. At the data level, using the U.S. CLA as an example, scores must be obtained for students' cognitive abilities in critical thinking, analytical reasoning, real-world problem-solving, and writing. The first three skills reflect students' capacity to evaluate and analyze information, primarily measured through dimensions such as "material analysis" and "drawing conclusions." For instance, the "drawing conclusions" dimension examines whether arguments are logical and data-rooted. Writing ability manifests the aforementioned three skills and students' logical thinking, evaluated through argument clarity and conciseness, structural effectiveness, writing quality, and content appeal [13-14]. During assessment, evaluators assign scores based on a four-level rating scale according to CLA scoring criteria. While standardized testing offers advantages in evaluating student abilities, its practical implementation faces challenges due to the requirement for scientifically validated test instruments and significant influence from factors such as student response attitudes, which may prevent results from accurately reflecting objective conditions [15].

Scale methods collect report data from students or teachers about evaluation objects to comprehensively understand actual conditions. Using student self-report scales as an example, these instruments focus on investigating student backgrounds, learning processes, and experiences, with collected feedback data serving as a basis for measuring value-added performance and identifying corrective issues [16]. Representative U.S. self-report scales include the College Student Experiences Questionnaire and the National Survey of Student Engagement, which reflect teacher or school performance. At the data level, using the UK's National Student Survey as an example, data primarily includes student satisfaction ratings across eight dimensions: curriculum teaching, learning opportunities, assessment and feedback, academic support, organization and management, and learning resources, reflecting school performance. Specifically, 27 questions such as "teachers explain problems well," "teachers make research topics lively and interesting," and "this course is very inspiring" measure student satisfaction [17]. In China, methods based on coursework examination scores are widely applied, primarily using students' scores in subjects such as Chinese, mathematics, English, or natural science from various examinations as data sources. Scores from midterm, final, entrance, and graduation examinations are collected and processed to evaluate relevant subjects' value-added contributions.

3.2 Value-Added Assessment Models and Methods

The following sections categorize and summarize current evaluation models used to assess value-added contributions of students, teachers, or schools based on different data acquisition methods.

3.2.1 Exam Score-Based Value-Added Assessment Models and Methods

(1) Mean and Standard Score Algorithms

Methods calculating value-added through mean or standard score differences directly measure value-added by score differences between two examinations. For instance, a high school in Hongkou District, Shanghai, evaluating teachers collected Chinese, mathematics, and English scores from two classes at the beginning and end of the term, calculating average scores for each subject in both examinations and using average score changes to measure teachers' impact on student progress. To reduce the influence of factors such as test difficulty and item differences on evaluation results, some scholars standardize scores during evaluation, using the difference between standardized beginning and end-of-term average scores to represent actual value-added. The standard score calculation method uses the sum of subject scores divided by the sum of full marks across subjects, multiplied by 100 [18-19]. Although mean and standard score difference methods are operationally simple and straightforward, they require sufficiently large sample sizes to avoid evaluation result distortion caused by substantial student score fluctuations.

(2) Value-Added Model Algorithms

The Value-Added Model (VAM) comprises linear models of varying complexity, with value-added primarily calculated from residuals between predicted and actual values obtained through linear models. This model category enjoys widespread use and high recognition in the value-added assessment field [20]. Zhang Lewei employed single-factor linear regression to calculate high school seniors' value-added, collecting data on Chinese, mathematics, and English scores from city-wide unified examinations one year apart. Using the two sets of scores, a linear estimation model for student performance was fitted. Based on students' initial scores, the model predicted their second-year examination scores, with the difference between actual and predicted second-year scores representing the value-added estimate. The value-added function is shown in Formula (1).

MATH_1 is the estimated value-added of student scores, MATH_2 is the student's actual score, and MATH_3 is the predicted score [21]. Some scholars have further considered student background characteristics, using random effects models to calculate predicted scores based on student background factors and subsequently computing value-added [22]. When actual scores (MATH_4) exceed expected scores (MATH_5), positive value-added is achieved, with the average of all students' value-added scores representing the school's educational value-added effect. The basic algorithm is shown in Formula (2).

Value added = average of MATH_6 for all students in a college Oi − Ei

The most widely applied model within VAM is the Hierarchical Linear Model (HLM), which can handle nested data structures, separate factors unrelated to evaluation objects, and reasonably assess actual value-added performance [23]. The basic approach involves designing multilevel linear regression equations based on selected variables that may influence evaluation results, then examining the impact of variables across dimensions to determine the optimal evaluation model, using residuals between predicted and actual results to estimate evaluation objects' value-added [24]. For example, when evaluating school value-added using a two-level HLM model, a student-level evaluation model must first be fitted, as shown in Formula (3).

MATH_7

where MATH_8 represents students, MATH_9 represents schools, MATH_{10} represents test scores, MATH_{11} represents entrance scores, MATH_{12} represents average scores, MATH_{13} represents school average scores, MATH_{14} indicates the first-level regression slope of student entrance academic ability scores, and MATH_{15} represents residuals. Second, a school-level regression model is fitted as shown in Formula (4).

MATH_{16}

where MATH_{17} represents school institutional characteristics and MATH_{18} represents the second-level regression slope of school characteristics. These formulas show that the total residual describing actual school value-added = MATH_{19} . Multilevel linear evaluation models that consider background factors to assess evaluation subjects' net value-added have also been widely applied in China [25-26].

(3) Student Growth Percentile Model

The Student Growth Percentile (SGP) model calculates a relative value-added measure, where "value-added" is assessed by the proportion of students exceeding peers at similar achievement levels. This method addresses evaluation challenges for high-achieving students with limited room for improvement and students with different starting points [27] and is currently widely used in the U.S. education evaluation system [28]. The specific calculation process involves first grouping students into different achievement levels based on existing test scores, then comparing changes in students' relative positions within their original achievement groups in subsequent examinations. This change reflects student progress, measured through an SGP rating scale ranging from 1-99, indicating the percentage of students in their achievement group that they have surpassed [29], with higher percentages indicating greater progress. The percentile regression model is shown in Formula (7).

MATH_{20}

where MATH_{21} represents a specific percentile, MATH_{22} represents retest scores, MATH_{23} represents initial test scores, and τ represents a piecewise linear loss function calculated as shown in Formula (8). Integrating these yields 99 quantile regression equations to calculate each student's SGP rating [30].

MATH_{24}

The Simplified Percentile Rank Growth Model is a simplified version of the SGP model. After collecting two sets of student test scores, the first test scores are sorted and each student's percentile rank score is calculated as shown in Formula (6).

MATH_{25}

where MATH_{26} is the ranking of a specific score in the overall score ranking, and MATH_{27} is the total frequency of that score occurrence, with student percentile rank scores ranging between 0-100. Subsequently, students at the same percentile rank in the first test are defined as a group, and their percentile rank within the group is recalculated using the second test results to represent student progress, with school value-added represented by the median of all students' growth percentile ranks [31].

To address estimation issues for nested data growth percentiles, Liu Yamei [32] and Zhou Yuan [33] incorporated control variables such as gender, social background, and household registration into the traditional SGP model, constructing nested multilevel linear model quantile regression mixture models to estimate student percentile ranks. In Zhou Yuan's research, variables such as parental occupation were pre-collected using scales as background factors. After collecting students' two test scores, a multilevel linear quantile regression model was constructed as shown in Formula (9), where MATH_{28} represents fixed effects and MATH_{29} represents random effects.

The first step in student-level quantile estimation involves using IRT models to estimate students' ability values in both tests, then using the second-year latent ability predicted value as the dependent variable and the first-year ability estimate as the independent variable to construct a multilevel linear quantile regression model. The difference between predicted and observed latent ability values at each percentile is calculated, with the percentile having the smallest absolute prediction error designated as the student's final growth percentile. This method can be considered an improvement on the SGP model, though its calculation process is relatively cumbersome and complex. Overall, compared with other value-added models, the SGP model offers the advantage of being less affected by extreme or outlier values, though it generally requires larger sample sizes; otherwise, the stability of evaluation ratings may be compromised [34].

3.2.2 Standardized Test-Based Value-Added Assessment Models and Methods

(1) Standard Score Difference and Average Value-Added Models

Regarding value-added quantification through standard score differences, Li Yi et al. used a standardized test covering four dimensions—reading interest, reading strategies, reading habits, and reading ability—to conduct pre- and post-tests of students' reading literacy, using differences between standardized test scores to quantify reading literacy value-added. Considering the impact of student background differences on evaluation results, they administered questionnaires to parents, teachers, and school administrators to obtain relevant background information and conducted more comprehensive and in-depth analyses of students' reading literacy value-added using hierarchical multiple regression models based on this background information [34]. In addition to using standard score differences, Wang Xiaohua employed an average value-added model, using average score differences in reading, mathematics, and science to quantify school impact on students in each subject. Multivariate linear regression models can also be used to construct separate multivariate linear regression models for the three subjects at the school level, calculating school value-added from differences between actual and predicted student scores [36].

(2) Hierarchical Linear Models
Hierarchical linear models represent the mainstream method used by scholars to evaluate student value-added using standardized test data. Li Jiahui et al. used a traditional hierarchical evaluation model—the gain score model—to evaluate students' listening learning effects. The value-added assessment model constructed with post-test scores as the dependent variable and initial test scores as the independent variable is shown in Formula (10).

MATH_{30}

where MATH_{31} represents students, MATH_{32} represents classes, MATH_{33} represents test occasions, MATH_{34} represents class means, MATH_{35} represents class random effects, MATH_{36} represents student random effects, and MATH_{37} represents the score of the MATH_{38} th test for the MATH_{39} th class. The difference between post-test and initial test scores divided by the initial test score serves as the criterion for evaluating student value-added [37]. Yang Zhiming [38], Shao Yueyang [39], and Hu Zhiqiao [40] further considered the impact of student and school variables on student achievement, using hierarchical linear regression models to conduct targeted tracking and evaluation of value-added in Chinese literacy, core competencies in Chinese, mathematics, and English, and students' knowledge and skill levels.

(3) Fixed Effects Models
Fixed effects models based on ELS standardized test data are frequently used in academic research [41-42], with the advantage of stripping away the influence of factors unrelated to evaluation objects to ensure objective results. For example, when evaluating principal value-added, to avoid confounding judgments about principals' and schools' fixed characteristics' impact on value-added, fixed effects models must be used to control for these characteristics when measuring principal value-added, as shown in Formula (11). This model uses ordinary least squares for value-added estimation [43].

MATH_{40}

where MATH_{41} is the score from one year prior, MATH_{42} is a vector of student characteristic variables, MATH_{43} is a vector of school-level demographic characteristics, MATH_{44} is a vector of time-varying principal-level characteristics, MATH_{45} represent principal, school, and year effects respectively, and MATH_{46} is the error term.

(4) Ordinary Least Squares Models Based on Cross-Sectional Data
The aforementioned methods evaluate value-added using longitudinal data from the same cohort. In contrast, organizations such as the American Association of State Colleges and Universities employ cross-sectional designs to evaluate educational effects in writing and critical thinking. Specifically, test scores from first-year and fourth-year students are collected separately, and ordinary least squares regression models are used to calculate predicted scores for both cohorts. Residuals between actual and predicted scores for both cohorts are then calculated, with school value-added represented by the difference between the two cohorts' score residuals. This method is relatively cost-effective and more feasible to implement [44], but it does not evaluate the same subjects and fails to account for nested data characteristics of student factors, resulting in potentially one-sided evaluation results [45].

(5) Rasch Model Based on IRT Item Response Theory
IRT item response theory evaluates relevant subjects' value-added based on changes in student ability, where "items" are test questions and "item responses" are students' answers to specific questions. Zhu Zhemin et al. tested sixth-grade graduates' mathematics academic quality twice and used the Rasch model from IRT methods to estimate value-added across various dimensions, using changes in student ability between the two tests to represent value-added. The Rasch model is shown in Formula (12).

MATH_{47}

where MATH_{48} represents student ability, and MATH_{49} represent discrimination parameters and difficulty parameters for the MATH_{50} th question, respectively. IRT's advantage lies in its ability to measure student ability levels through responses to each test item, with ability changes representing student progress, thereby overcoming the drawbacks of "score-only" evaluation models. This method is also an ideal tool for score equating. IRT-based academic quality evaluation methods are gradually gaining academic attention [46-47].

3.2.3 Scale Data-Based Value-Added Calculation Methods

Among value-added assessment methods based on scale data, the simplest category calculates value-added using score differences after conversion to a 100-point scale. For example, Pan Jiening et al. used the Problem-Solving Ability Scale, which includes four dimensions such as understanding and analyzing problems and problem-solving ability, to collect student test data twice. They calculated value-added in problem-solving ability for vocational college students using differences between pre- and post-test scores converted to a 100-point scale [48]. Some scholars have adopted a knowledge graph value-added assessment model to visualize teacher value-added. Specifically, teacher evaluation questionnaires containing student knowledge, ability, and competency elements are distributed before and after class for students to complete via a cloud platform. The data is then imported into a Neo4j database to generate student learning graph portraits, and the PageRank algorithm is used to assign values to graph nodes. Weighted scores of student learning graphs are calculated using node weight values and expressed as percentages, with the resulting percentage representing the value-added score. The calculation formula is shown as (13).

MATH_{51}

where MATH_{52} represents the weighted score of the student graph, MATH_{53} represents the weight value of node number 1 in the reference graph, and MATH_{54} represents the weight value of node number 1 in the learning graph, with n representing the number of learning graph nodes and m representing the number of reference graph nodes [49]. In addition to these models, traditional hierarchical linear models are also fully applied in processing scale data to calculate value-added scores, with specific evaluation processes and methods essentially consistent with those described previously [50].

3.2.4 Comprehensive Data-Based Value-Added Calculation Methods

Comprehensive data-based calculation methods utilize data generated by multiple subjects and collected through multiple channels. For example, Zhang Liang obtained evaluation data through examination scores and student and teacher questionnaires, measuring school education value-added by the difference between individual student score change averages and all student score change averages [51]. Bacher-Hicks A et al. used standardized state test scores, project test scores, and rating scale scores from student evaluations of teachers. After standardizing these data, they used ordinary least squares regression equations to obtain teacher annual residuals as the basis for value-added assessment [11], with the regression formula shown as (14).

MATH_{55}

where MATH_{56} represents the standardized state test scores of students taught by the teacher in the academic year, MATH_{57} is the student baseline score, MATH_{58} is the vector of average characteristics of peers in the same class, MATH_{59} is the vector describing student background information, MATH_{60} is the grade-year fixed effect, MATH_{61} is the fixed effect, and MATH_{62} is the student-level error. Teacher value-added is estimated by the annual residual represented by the average of student-level errors, with similar calculations yielding teacher annual residuals based on project test scores and scale scores. Comprehensive teacher value-added estimates are obtained by integrating these rating data. Lee J used standardized test scores and teacher rating scale data to evaluate student value-added from both academic and social-emotional skill development perspectives. Students' academic skills were assessed through standardized tests in reading, mathematics, and science, while social-emotional skill dimension value-added was calculated from teacher ratings of students' learning methods, self-control, and interpersonal skills. The study also collected student background factors and constructed a three-level HLM model including measurement, student, and school levels, estimating school value-added effects by the average difference between students' actual and predicted growth values [23].

Although abundant research results have been achieved in value-added assessment both domestically and internationally, not all value-added assessment models produce accurate results [52], and model selection also influences evaluation outcomes [36]. This paper systematically reviews the concepts, measurement models, and measurement methods of value-added assessment within the context of China's new-era education evaluation reform, providing references for the reform and practice of higher education evaluation and graduate education quality assessment in China.

References

[1] Central Committee of the Communist Party of China, State Council. Overall Plan for Deepening the Reform of Education Evaluation in the New Era[EB/OL]. (2020-10-13)[2023-09-28]. https://www.gov.cn/zhengce/2020-10/13/content_{5551032}.htm.

[2] Coleman J S, Campbell E Q, Hobson C J, et al. Equality of educational opportunity[M]. Washington, DC: US Department of Health, Education & Welfare, 1966.

[3] Ren Chunrong. Value-added measurement: A scientific approach to fairly using test scores to evaluate school effectiveness[J]. China Examinations, 2007(4):12-16+38.

[4] Jiang Hua, Wang Miao. Research progress on value-added assessment in higher education[J]. Teaching Research, 2022, 45(5):8-15+12.

[5] William L, Sanders S P H. Research Findings from the Tennessee Value-Added Assessment System (TVAAS) Database: Implications for Educational Evaluation and Research[J]. Journal of Personnel Evaluation in Education, 1998, 12(3):247-256.

[6] Heck R H. Examining the Relationship Between Teacher Quality as an Organizational Property of Schools and Students' Achievement and Growth Rates[J]. Educational Administration Quarterly, 2016, 43(4):399-432.

[7] Bian Yufang, Sun Liping. Progress in teacher value-added assessment and recommendations for its application in China[J]. Teacher Education Research, 2015, 27(1):88-95+112.

[8] Qiu Junping, Wen Tingxiao. Evaluation Science: Theory · Method · Practice[M]. Beijing: Science Press, 2010.

[9] Guan Dandan, Han Ning, Zhang Jianshi. Deepening college entrance examination evaluation reform based on "four evaluations" and serving "five types of subjects"[J]. China Examinations, 2021(3):1-8.

[10] Leckie G, Goldstein H. The importance of adjusting for pupil background in school value-added models: A study of Progress 8 and school accountability

in England[J]. British Educational Research Journal, 2019, 45(3):518-537.
[11] Bacher-Hicks A, Chin M J, Kane T J, et al. An experimental evaluation of three teacher quality measures: Value-added, classroom observations, and student surveys[J]. Economics of Education Review, 2019, 73:1-15.
[12] PAPAY J P. Refocusing the debate: Assessing the purposes and tools of teacher evaluation[J]. Harvard Educational Review, 2012, 82(1):123-141.
[13] Mongkuo M, Lucas N, Walsh K. Initial Validation of Collegiate Learning Assessment Performance Task Diagnostic Instrument for Historically Black Colleges and Universities[J]. The British Journal for the Philosophy of Science, 2016, 928-950.
[14] Hersh R H. Collegiate learning assessment (CLA): Defining critical thinking, analytical reasoning, problem solving, and writing skills[EB/OL]. (2007)[2022-11-10]. http://www.teaglefoundation.org/learning/pdf/hersh_{ctdefinitions}.pdf.
[15] Zhou Tingyong, Du Ruijun, Zhang Xieyu. Analysis and research on standardized assessment tools for American college student learning outcomes[J]. Fudan Education Forum, 2014, 12(5):84-90.
[16] Kane D, Williams J, Cappuccini-Ansfield G. Student Satisfaction Surveys: The Value in Taking an Historical Perspective[J]. Quality in Higher Education, 2008, 14(2):135-155.
[17] Sofroniou A, Premnath B, Poutos K. Capturing Student Satisfaction: A Case Study on the National Student Survey Results to Identify the Needs of Students in STEM Related Courses for a Better Learning Experience[J]. Education Sciences, 2020, 10(12):1-22.
[18] Wang Binhua. Teacher evaluation: Value-added assessment method[J]. Education Theory and Practice, 2005(23):22-25.
[19] Cao Jian, Hu Shankang, Li Qinglong. Application of two value-added assessment methods in basic education evaluation[J]. Educational Measurement and Evaluation, 2010(8):4-7.
[20] Lopez-Martin E, Kuosmanen T, Gaviria J L. Linear and nonlinear growth models for value-added assessment: an application to Spanish primary and secondary schools' progress in reading comprehension[J]. Educational Assessment, Evaluation and Accountability, 2014, 26(4):361-391.
[21] Zhang Lewei. Research on value-added assessment of middle school students[D]. Tianjin: Tianjin Normal University, 2012.
[22] Shavelson R J, Domingue B W, Mariño J P, et al. On the practices and challenges of measuring higher education value added: the case of Colombia[J]. Assessment & Evaluation in Higher Education, 2016, 41(5):695-720.
[23] Lee J, Kim T, Su M. Reassessing school effectiveness: Multi-objective value-added measures (MOVAM) of academic and socioemotional learning[J]. Studies in Educational Evaluation, 2021, 68:1-11.
[24] Kong S, Jiang L, Li G, et al. Value-added Evaluation of College Students' English learning efficiency based on Hierarchical linear model[C]. 2014 International Conference on Advanced ICT (ICAICTE-2014). Atlantis Press, 2014, 116-120.
[25] Bian Yufang, Lin Zhihong. Value-added assessment: A school evaluation model under the concept of green admission rate[J]. Journal of Beijing Normal

University, 2007(6):11-18.
[26] Chen Anqi, Guan Dandan. Comparative analysis and empirical research of several value-added assessment methods[J]. China Examinations, 2022(9):54-62.
[27] Yang Zhiming, Peng Liyi, Li Yang. Ceiling effects in value-added assessment and approaches to address them[J]. Educational Measurement and Evaluation, 2020(12):3-7.
[28] Kurtz M D. Value-Added and Student Growth Percentile Models: What Drives Differences in Estimated Classroom Effects?[J]. Statistics and Public Policy, 2018, 5(1):1-8.
[29] Wang Shuaiming, Yao Shoumei, Wang Meng, et al. Empirical research on value-added assessment based on the student growth percentile model[J]. China Examinations, 2020(9):46-51.
[30] Xie Xiaoqing. Student growth percentile model for growth assessment: Experience from the United States[J]. Educational Measurement and Evaluation, 2019(6):3-9.
[31] Xu Luming. Academic value-added assessment based on the simplified percentile rank growth model[J]. China Examinations, 2021(3):9-14.
[32] Liu Yamei. Empirical research on value-added assessment of junior high school mathematics based on regional data[J]. Shanghai Educational Research, 2022(8):34-39.
[33] Zhou Yuan, Liu Hongyun. Analysis of estimation methods for growth percentiles of nested data in educational value-added assessment: Application of multilevel linear quantile regression models[J]. China Examinations, 2020(9):32-39.
[34] Betebenner D. Norm-and Criterion-Referenced Student Growth[J]. Educational Measurement: Issues and Practice, 2009, 28(4):42-51.
[35] Li Yi, Zheng Pengyu, Huang Yiming. Value-added assessment in basic education: Connotation, value, and empirical research—Taking reading literacy of primary and secondary students in the new era as an example[J]. China Educational Technology, 2022(10):47-55+71.
[36] Wang Xiaohua. School value-added assessment research based on PISA data[J]. Research in Educational Development, 2021, 41(10):8-17.
[37] Li Jiahui, Hu Zhaoyang. Construction and application of a listening learning effect (dynamic) value-added assessment system[J]. Examination and Evaluation, 2020(3):14-21.
[38] Yang Zhiming, Zhang Feng. Application of hierarchical linear regression methods in value-added assessment[J]. Educational Measurement and Evaluation, 2022(2):3-11.
[39] Shao Yueyang, Liu Jian. Value-added assessment: Focusing on schools' support for every student's growth—An empirical study based on data from a district in Beijing[J]. China Examinations, 2020(9):40-45.
[40] Hu Zhiqiao, Zhong Yang, Wang Jing. Construction and practice of a value-added assessment model for primary and secondary students' academic levels—Based on Guangzhou's compulsory education stage sunshine evaluation research[J]. Education Theory and Practice, 2022, 42(11):18-22.

[41] Isenberg E, Max J, Gleason P, Deutsch J. Do Low-Income Students Have Equal Access to Effective Teachers?[J]. Educational Evaluation and Policy Analysis, 2022, 44(2):234-256.
[42] Bitler M, Corcoran S P, Domina T, et al. Teacher Effects on Student Achievement and Height: A Cautionary Tale[J]. Journal of Research on Educational Effectiveness, 2021, 14(4):900-924.
[43] Dhuey E, Smith J. How school principals influence student learning[J]. Empirical Economics, 2018, 54(2):851-882.
[44] Voluntary System of Accountability. Background on learning outcomes measures[EB/OL]. (2008)[2023-11-10]. www.voluntarysystem.org/index.cfm?page=about_{cp}.
[45] Liu O L. Value-added assessment in higher education: a comparison of two methods[J]. Higher Education, 2010, 61(4):445-461.
[46] Zhu Zhemin, Kong Xiangyi. Value-added assessment and application of mathematics academic quality based on IRT[J]. Educational Measurement and Evaluation, 2022(5):21-31.
[47] Yang Zhiming, Jia Lixin, Lü Longmei. Value-added assessment of reading ability based on IRT equating[J]. Educational Measurement and Evaluation, 2018(10):5-12.
[48] Pan Jiening. Exploration and practice of value-added assessment of problem-solving ability in vocational college students—An empirical analysis based on assessments from 6 vocational colleges[J]. Vocational and Technical Education, 2022, 43(20):54-61.
[49] Zhang Ningyu. How to evaluate the effectiveness of vocational school teachers' teaching—A strategy based on the student knowledge graph value-added model[J]. Chinese Vocational and Technical Education, 2022(17):49-54.
[50] Du Ping, Yang Zhongchao. Value-added assessment of rural junior high school effectiveness—An empirical analysis based on survey data from five western provinces in China[J]. Journal of Beijing Normal University (Social Sciences Edition), 2011(6):91-97.
[51] Zhang Liang. Research on value-added assessment of ordinary high school students[D]. Jinan: Shandong Normal University, 2010.
[52] Amrein-Beardsley A, Geiger T. Methodological Concerns About the Education Value-Added Assessment System (EVAAS): Validity, Reliability, and Bias[J]. SAGE Open, 2020, 10(2):1-15.

Note: Figure translations are in progress. See original paper for figures.