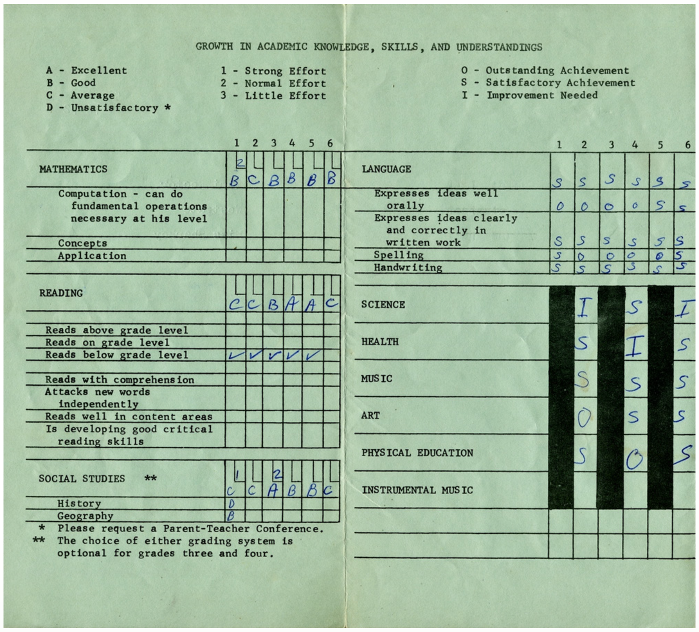
NEPC Review: Making the Grade: Accounting for Course Selection in High School Transcripts With Item Response Theory (Annenberg Institute for School Reform at Brown University, December 2024)



GROWTH IN ACADEMIC KNOWLEDGE, SKILLS, AND UNDERSTANDINGS

A - Excellent
B - Good
C - Average
D - Unsatisfactory *

1 - Strong Effort
2 - Normal Effort
3 - Little Effort

O - Outstanding Achievement
S - Satisfactory Achievement
I - Improvement Needed

	1	2	3	4	5	6
MATHEMATICS	2 B	C	B	B	B	B
Computation - can do fundamental operations necessary at his level						
Concepts						
Application						
READING	C	C	B	A	A	C
Reads above grade level						
Reads on grade level						
Reads below grade level	✓	✓	✓	✓	✓	
Reads with comprehension						
Attacks new words independently						
Reads well in content areas						
Is developing good critical reading skills						
SOCIAL STUDIES **	1 C	C	2 A	B	B	C
History	D					
Geography	B					

* Please request a Parent-Teacher Conference.
** The choice of either grading system is optional for grades three and four.

	1	2	3	4	5	6
LANGUAGE	S	S	S	S	S	S
Expresses ideas well orally	O	O	O	O	S	S
Expresses ideas clearly and correctly in written work	S	S	S	S	S	S
Spelling	S	O	O	O	O	S
Handwriting	S	S	S	S	S	S
SCIENCE		I		S		I
HEALTH		S		I		S
MUSIC		S		S		S
ART		O		S		S
PHYSICAL EDUCATION		S		O		S
INSTRUMENTAL MUSIC						

Joy Rector/Shutterstock.com

Reviewed by:

Michael Russell
Boston College

March 2025

National Education Policy Center

School of Education
University of Colorado Boulder
nepc.colorado.edu

Acknowledgements

NEPC Staff

Faith Boninger
Publications Manager

Patricia Hinchey
Academic Editor

Elaine Duggan
Production Design

Alex Molnar
NEPC Director

Kevin Welner
NEPC Director

Suggested Citation: Russell, M. (2025). *NEPC Review: Making the grade: Accounting for course selection in high school transcripts with item response theory*. Boulder, CO: National Education Policy Center. Retrieved [date] from https://nepc.colorado.edu/review/gpa

Funding: This review was made possible in part by funding from the Great Lakes Center for Educational Research and Practice.

The National Education Policy Center (NEPC), a university research center housed at the University of Colorado Boulder School of Education, sponsors research, produces policy briefs, and publishes expert third-party reviews of think tank reports. NEPC publications are written in accessible language and are intended for a broad audience that includes academic experts, policymakers, the media, and the general public. Our mission is to provide high-quality information in support of democratic deliberation about education policy. We are guided by the belief that the democratic governance of public education is strengthened when policies are based on sound evidence and support a multiracial society that is inclusive, kind, and just. Visit us at: http://nepc.colorado.edu

NEPC
NATIONAL EDUCATION POLICY CENTER

NEPC Review: Making the Grade: Accounting for Course Selection in High School Transcripts with Item Response Theory (Annenberg Institute for School Reform at Brown University, December 2024)

Reviewed by:

Michael Russell
Boston College

March 2025

Summary

Use of test scores to inform college decisions is hotly debated. Absent SAT scores, GPA serves as the primary indicator of student readiness for college. A student's GPA, however, is influenced by the difficulty of courses taken. An Annenberg Institute report explores the creation of a new measure, termed Transcript Strength, that seeks to adjust a student's GPA based on the difficulty of courses completed. To create this new measure, the report treats courses like questions on an educational test. It then generates a measure of transcript strength using a common technique employed in measuring academic achievement, a partial credit model based on Item Response Theory. The report provides preliminary evidence that the new tool provides information about student high school achievement that differs from both GPA and the SAT. Although there are several ways the report's analyses could be improved and some potential barriers to the tool's widespread use, its presentation is sound and reasonable. As the report itself concludes, the measure holds promise from a theoretical perspective to be a more informative indicator of high school achievement than GPA, but it is not yet ready for implementation. To be clear, the measure of transcript strength is not yet ready for use by policymakers or admission officers. Instead, the approach needs further research and development.

NEPC
NATIONAL EDUCATION
POLICY CENTER

NEPC Review: Making the Grade: Accounting for Course Selection in High School Transcripts with Item Response Theory (Annenberg Institute for School Reform at Brown University, December 2024)

Reviewed by:
Michael Russell
Boston College

March 2025

I. Introduction

For several decades, observers have raised concerns about the use of the SAT and ACT for college admissions.[1] Recently, several colleges and universities have stopped using test scores.[2] Although some observers argue this change makes the admission process more equitable, others argue reliance on other sources of evidence, such as GPA, personal statements, and letters of recommendation, contribute to similar or worse inequities in admissions.[3]

Despite this disagreement, there is general consensus that GPA is compromised by differences in course taking patterns, differences in curriculum among schools, and differences in the difficulty of courses. In a recent report, *Making the Grade: Accounting for Course Selection in High School Transcripts with Item Response Theory,* published by the Annenberg Institute, Kenneth Shores and Sanford Student describe an effort to address some of these concerns by developing a new measure of student high school achievement that focuses on *transcript strength*.[4] To do so, the report applies a measurement technique to adjust a student's GPA based on the difficulty of courses completed.[5]

II. Findings and Conclusions of the Report

The report demonstrates the feasibility of applying a measurement technique in a new way to assess transcript strength and presents evidence that it improves on GPA. The new tool distinguishes between students who have the same GPA but completed courses that differ in difficulty, providing higher transcript strength scores to students who completed more difficult courses. The report also finds that the detection of differences between subgroups of students differs depending on whether one examines the new measure or the SAT.

Finally, some evidence is presented that the new measure reflects the construct termed *transcript strength* and provides stronger predictions of some college outcomes. Although the report acknowledges that additional research is needed before transcript strength is used as part of the higher education admission decision process, the report concludes this new measure improves on GPA by incorporating information about course difficulty into the measurement process.

III. The Report's Rationale for Its Findings and Conclusions

The report presents a variety of analyses to examine characteristics of the transcript strength measure and to compare it to GPA and SAT. This section focuses on results from three analytic categories.

The first set of analyses provides evidence that the new measure reflects transcript difficulty. The report shows that the estimated difficulty of courses aligns with conventional understanding of easier and harder courses. For example, Advanced Placement courses were generally considered more difficult than standard courses, and STEM courses were typically considered more difficult than humanity, arts, and physical education courses. In addition, the report indicates that higher ability students typically enroll in more difficult courses and, although this pattern tends to deflate the GPA of higher achieving students relative to lower achieving students, the transcript strength measure distinguishes between these subgroups of students.

The second set of analyses provides evidence the transcript strength measure has stronger predictive power of college outcomes than the SAT or GPA. The report finds it is a stronger predictor of college completion and earnings after completing college. The report also finds that students with higher transcript strength scores tend to attend more selective schools.[6]

The third set of analyses compares transcript strength and SAT scores between various subgroups of students. These analyses find that whereas there is almost no difference in SAT scores between male and female students, female students have higher transcript scores than males. In contrast, the difference in SAT and transcript strength scores is similar between socioeconomic status groups and for students who are White and either Asian or Black. For students who are Hispanic, analyses reveal a smaller difference in transcript strength scores versus SAT. These analyses indicate the choice of metric can impact the detection of educational inequities.

IV. The Report's Use of Research Literature

To situate the study, the report draws on literature that considers the fairness of college admission tests and the potential impact on equity that occurs when test scores are not used during the admission process. The report also draws on literature to document various ways that GPA is used as an indicator of academic achievement in admissions and quantitative research. Through this review, the report notes some limitations to GPA that result from differences in student course-taking patterns and difficulty of courses.

To inform the technique used to develop the measure of transcript strength, the report references literature on Item Response Theory (IRT). The report considers several IRT-based techniques commonly used to produce scores for tests that contain open-response items (essay questions, for example) that can receive partial credit.[7] Based on this literature, the report selects a model that aligns with the measurement aims and the data available for analysis. Noting this is the first known attempt to apply IRT to create a measure of transcript strength, the report draws on three similar applications of IRT to transcript data to inform this novel application.

Although the literature base is sufficient for situating the study and for justifying the measurement techniques, there are two bodies of literature that are underutilized. First, when considering future research, the report notes that differences in the difficulty of courses that occur across schools and between educators are not accounted for in the current approach. To explore the impact such differences may have on the measure of transcript strength, the report suggests using techniques that treat these differences as random factors. The report does not consider, however, other IRT-based approaches that adjust a measure based on these effects.[8]

The second underutilized body of literature focuses on validity. The report does present some evidence that supports the new measure's validity. However, over the past century, a large body of literature on validity has been produced and the concept has evolved considerably, with definitions generally shifting from a focus on tests or measures themselves to a focus on the interpretation and uses of their results.[9] Although the report references the most recent version of *Standards for Educational and Psychological Testing*,[10] which reflects the current conception, it nevertheless relies on dated notions of discrete forms of *construct validity* (does the test measure what it says it measures?) and *predictive validity* (does the tool predict what it claims to predict?) to justify the new measure.

V. Review of the Report's Methods

The report works to demonstrate: that the effort and strategy to create a measure of transcript strength is reasonable; that it can and does measure a concept that is not directly measurable (employing the earlier conception of *construct validity*); and that it can and does predict what it claims to predict (employing the earlier conception of *predictive validity*).

Integral to the proposed measurement is the assumption that high school courses function similarly to test questions. Just as test questions differ in difficulty, with some being relatively easy to answer correctly and others harder, the report recognizes that high school courses differ in difficulty, making it easier to receive a high grade in some courses and harder in other courses. Applying this reasoning, the report treats the courses taken by a student as if they were a set of test items. However, because students take different sets of courses, with some sets being more challenging than others, the report treats a student's high school course-taking experience similarly to an adaptive test.[11]

In an adaptive test, a large pool of items is developed. Different subsets of items from this pool are then administered to different students. Typically, the items administered to a given student are adapted based on how well the student has performed on previous items. Students who respond correctly to relatively easy items are presented with items that increase in difficulty while those who respond incorrectly are presented with easier items. This process is repeated until the student begins to answer more difficult items incorrectly or vice versa.

Whereas the selection of items for an adaptive test is driven by an algorithm that operates in the background, the report assumes that students similarly select courses based on their performance in prior courses and the perceived difficulty of future courses. Although this reasoning may hold in some cases, no evidence is presented to support this assumption and it seems reasonable that other factors, such as limited space in courses, parental influence, teacher guidance, and undergraduate education ambitions, may also influence course selection. Regardless, the core assumption that courses vary in difficulty and that a set of courses can be treated similarly to an adaptive test is reasonable.

To demonstrate the potential of the proposed measure, the report uses data collected from Delaware public schools.[12] The report treats courses with the same name as if they are the same test item. Although this assumption is practical for an exploratory study, it is unlikely to hold in reality. As the report observes, differences in curricular materials, educator expectations, and content coverage produce differences in the difficulty of courses offered across schools. Regardless, the report applies the measurement technique to simultaneously estimate the difficulty of courses, and the strength of each student's transcript based on the courses taken and the student's grades in those courses.

Although the way in which the measurement technique is applied is sound, the term *transcript strength* is open to question. Although the term implies that the scores produced by the tool measure the difficulty of the set of courses that form the transcript, in effect the resulting score is more analogous to a GPA that is adjusted based on the difficulty of courses the student completed. For this reason, it seems more appropriate to interpret the metric as an adjusted GPA. Moreover, there are places in the report that refer to the measure representing "student ability," "college readiness," and "student achievement."[13] Future development of the measure should include greater clarity about what trait the measure represents.

As noted above, the report uses dated conceptions of construct and predictive validity. Although the approaches used to examine these two types of validity are informative, they could be strengthened. While it is informative to compare the measure of course difficulty the new tool produces with conventional understanding of course difficulty, the report could use a more rigorous method to establish conventional understanding. In fact, the report does not reference any prior research on or describe any approach to establish conventional understanding.[14] A sounder approach might assemble a sample of educators and/or others who are familiar with variation in the content and difficulty of courses to rate or rank courses based on perceived difficulty.[15]

The report also seeks to demonstrate construct validity by comparing transcript strength with SAT scores using conventional methodology. This analysis indicates that the two measures have a strong correlation and are predictive of each other. It is unclear, however, how the comparison provides evidence that the new tool measures what it purports to measure, given that one focuses on transcript strength and the other on college readiness. As discussed above, this misalignment may be a product of the report's claim that the measure represents transcript strength rather than high school performance.

Finally, the report's analyses of predictive validity provide some evidence that the transcript strength measure is a stronger predictor than SAT for some college outcomes. Although these findings are promising, the SAT is designed to predict performance only during freshman year.[16] If one aim of predictive validity is to compare the measure of transcript strength with the SAT, an analysis that focuses on predicting freshman GPA would align more strongly with the intended purpose of the SAT.

Collectively, the analyses presented provide interesting insight into the relationship between transcript strength and GPA, SAT and characteristics of students.[17] However, it is unclear what the findings say collectively about the validity of the proposed transcript strength measure. Had the report applied a modern perspective, findings from these analyses might have contributed to a stronger and more coherent argument.

VI. Review of the Validity of the Findings and Conclusions

Given that the report acknowledges this is an exploratory study, the evidence presented is sufficient to support further research on the development and potential use of transcript strength. Based on the evidence provided, it seems the measure provides information that differs from GPA and SAT in meaningful and useful ways. It is unclear, however, what exactly the new measure is measuring. It seems to be measuring more than the strength of a student's transcript and may be more akin to a measure of high school performance given the difficulty of courses completed. If so, the report's conclusion that further research is both warranted and necessary prior to actual use of the measure during the college admission process is sound and reasonable.

VII. Usefulness of the Report for Guidance of Policy and Practice

As an initial step in developing a new measure of student achievement, the report offers promise for policymakers and college admission officers—although potential challenges remain. While the findings provide preliminary evidence regarding the measure's validity, a more robust, comprehensive, and modern approach to validation is needed. As part of this approach, evidence is necessary regarding potential uses of the measure and their consequences, particularly on different subgroups of students. As the report concludes, the measure holds promise from a theoretical perspective, but it is not yet ready for implementation.

Notes and References

1 Crouse, J. & Trusheim, D. (1988). *The case against the SAT*. University of Chicago Press.

Sandel, M.J. (2020). *The tyranny of merit: What's become of the common good*. Farrar, Straus, and Giroux.

Tough, P. (2019). *The years that matter most*. Random House.

2 Rather than stopping the use of test scores, some schools have made test scores optional.

Camara, W. (2024, Winter). Admission testing in higher education: Changing landscape and outcomes from test-optional policies. *Educational Measurement: Issues and Practice*, *43*(4), 104-111. Retrieved February 25, 2025, from https://doi.org/10.1111/emip.12651

Fair Test. (2025, February). *Test optional and test free colleges*. Retrieved February 24, 2025, from https://fairtest.org/test-optional-list/

3 Gooch, R.M., Belur, V.K., Haviland, S.B., & Liu, O.L. (2024, July). Test-optional policies: Impacts to date and recommendations for equity in admissions. *Journal of Postsecondary Student Success*, *3*(4), 1-19. Retrieved February 24, 2025, from https://files.eric.ed.gov/fulltext/EJ1432877.pdf

Paris, J.H., Torsney, B., Fiorot, S., & Pressimone Beckowski, C. (2022, July). The impact of optional: Investigating the effects of test-optional admissions policies. *Journal of College Access*, *7*(2), 7-29. Retrieved February 24, 2025, from https://files.eric.ed.gov/fulltext/EJ1372848.pdf

Rothstein, J. (2022, August). Qualitative information in undergraduate admissions: A pilot study of letters of recommendation. *Economics of Education Review*, *89*, 102285. Retrieved February 24, 2025, from https://doi.org/10.1016/j.econedurev.2022.102285

4 Shores, K.A. & Student, S.R. (2024, December). *Making the grade: Accounting for course selection in high school transcripts with Item Response Theory*, EdWorkingPaper No. 24-1009. Annenberg Institute, Brown University.

5 The report uses software, called *mirt*, to apply an Item Response Theory measurement technique called the Partial Credit Model. For ease of reference, throughout this review, the term *measurement technique* is used to refer to the Item Response Theory-based Partial Credit Model.

6 For expected earnings, the report finds that for each standard deviation increase in transcript strength, there is an approximately $7,100 increase in expected earnings, whereas the same improvement in GPA or SAT is associated with earning increases of approximately $6200 and $5000-6000 respectively. Separate analyses also show that transcript difficulty remains a strong predictor of college outcomes compared to the SAT even after controlling for GPA.

7 Partial credit items often receive scores that range from 0-2, 0-3, and so on depending on the scoring rules for the test item.

8 Linacre, J.M. & Wright, B.D. (2002, September). Construction of measures from many-facet data. *Journal of Applied Measurement*, *3*(4), 486-512.

9 Currently, validity is understood as a comprehensive body of evidence that focuses on technical properties of a measure, the ways in which measures are interpreted and used, as well as the consequences that result from those uses.

Sireci, S.G. (2009). Packing and unpacking sources of validity evidence: History repeats itself again. In R.W. Lissitz (Ed.), *The concept of validity: Revisions, new directions and applications, October 2008*. IAP Information Age Publishing.

Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York, NY: American Council on Education and Macmillan.

Kane, M.T. (1992, November). An argument-based approach to validity. *Psychological Bulletin, 112(*3), 527-35. Retrieved February 21, 2020, from https://psycnet.apa.org/buy/1993-11938-001

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

10 American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

11 Wainer, H., Dorans, N.J., Flaugher, R., Green, B.F., & Mislevy, R.J. (2000). *Computerized adaptive testing: A primer*. Routledge.

12 The data set contained grades for all courses taken by students who graduated from public schools between 2017 and 2021.

13 See pages 1, 9, 18, and 28, respectively, in Shores, K.A. & Student, S.R. (2024, December). *Making the grade: Accounting for course selection in high school transcripts with Item Response Theory*, EdWorkingPaper No. 24-1009. Annenberg Institute, Brown University.

14 Based on the information provided in the report, conventional understanding appears to be the understanding of those who authored the report.

15 This empirical evidence of perceived course difficulty could then be compared to the difficulty measures produced by the measurement technique to examine their alignment.

16 Camara, W.J. & Echternacht, G. (2000, July). *The SAT® I and high school grades: Utility in predicting success in college*. Research Notes. Retrieved February 24, 2025, from https://web.archive.org/web/20090106122457/http://www.collegeboard.com/research/pdf/rn10_10755.pdf

17 In addition to the analyses described in this section, the report conducts three additional sets of analyses: one that focuses on robustness to grade inflation; another that employs simulation methods to examine the accuracy with which the IRT-based measurement technique functions for various course taking patterns; and a third that compares differences in measures of transcript strength and SAT scores between subgroups of students. Although all three sets of analyses are sound, it is unclear how they relate to either construct validity or predictive validity, or more generally how the findings support a claim regarding the validity of the transcript strength measure.