Westat

Improving Lives Through Research®

Using Machine Learning to Improve Qualitative Coding of High School Courses to Enhance Data Quality

Judy H. Tang, Westat
Tom Krenzke, Westat
Jin Hui Xu, Westat
Karen Lo, Westat

2026 FedCASIC Virtual Workshop
Tuesday, April 21, 2026, 1:00 – 2:30 PM

Agenda

Context & Challenges

- NAEP HSTS Data and Coding Context
- Challenges in Qualitative Course Coding

Approach & Design

- Machine Learning Approach for Course Coding
- Multi-Phase Experimental Design

Results & Implications

- Results: Accuracy, Efficiency, and Consistency
- Implications for Data Quality in Federal Surveys

NAEP HSTS Data Collection and Processing Context

- National Assessment of Education Progress (NAEP) High School Transcript Study (HSTS) is a national data collection conducted by the National Center on Education Statistics (NCES) that focuses on understanding students' coursework and academic outcomes at the end of high school.

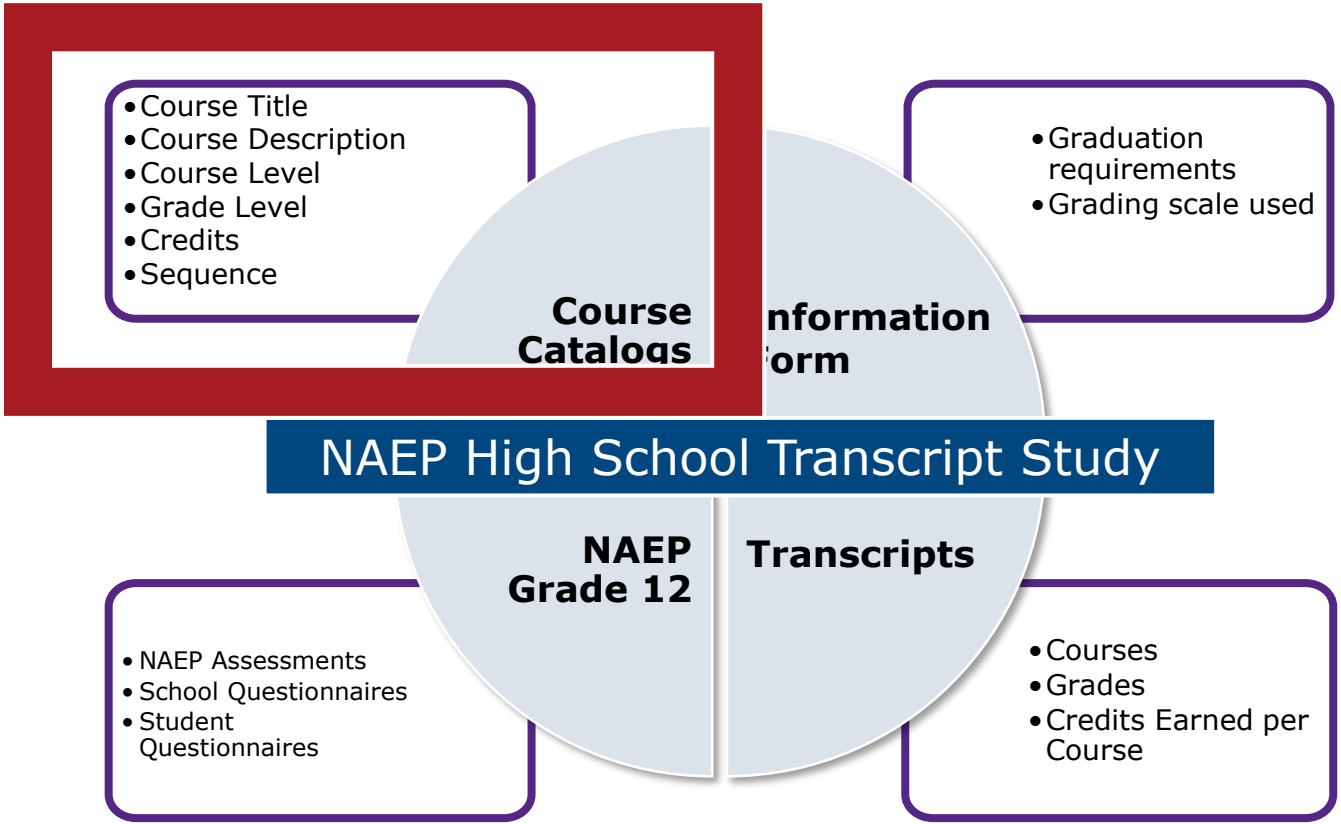
Selected a nationally representative sample of HS graduates.

Collected course catalogs, transcripts, and information forms from states, districts, and schools.

Coded courses from catalogs using the SCED classification system.

Standardized credits to Carnegie units and grades to a four-point scale.

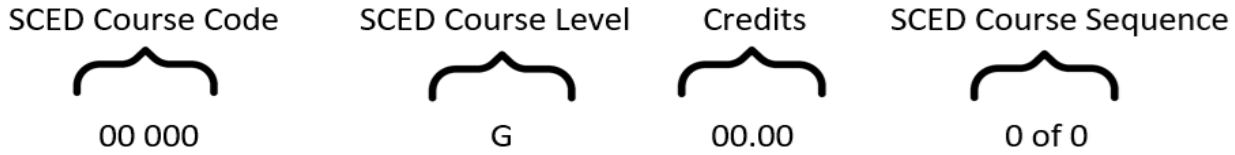
Data Components of NAEP HSTS



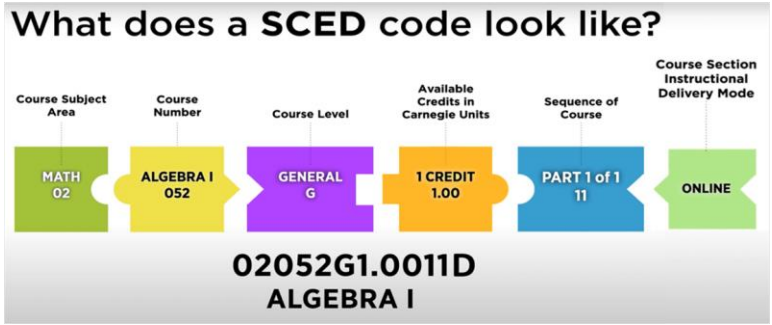
School Courses for the Exchange of Data (SCED) Coding System

- A voluntary classification system designed for prior-to-secondary and secondary education courses.
- Can be used to compare course information, maintain longitudinal data about student coursework, and efficiently exchange coursetaking records.
- Developed by the National Forum on Education Statistics through the NCES.
- States like Virginia and Iowa utilize SCED in their student data systems by using it as course identification codes.
- Other states and local education agencies have used portions of SCED or its structural framework for their own course IDs.

SCED Components and Structure



- The first component is a five-digit course code that identifies the course.
- The first two numbers of the base code identify the subject area.
- HSTS also includes flags for special education, course language taught, and online courses.

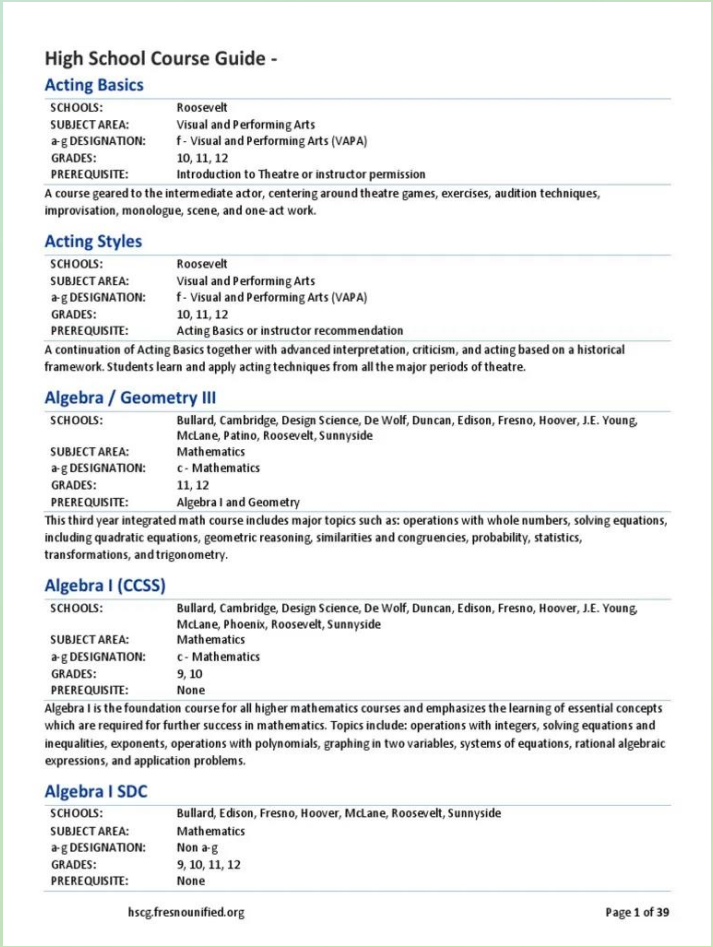


Source: National Forum on Education Statistics, National Center for Education Statistics (NCES).

- **23 Subject areas:**
  - 01: English Language and Literature
  - 02: Mathematics
  - 03: Life and Physical Sciences
  - 04: Social Sciences and History
  - 05: Visual and Performing Arts
  - 07: Religious Education and Theology
  - 08: Physical, Health, and Safety Education
  - 09: Military Science
  - 10: Information Technology
  - 11: Communications and Audio/Visual Technology
  - 12: Business and Marketing
  - 13: Manufacturing
  - 14: Health Care Sciences
  - 15: Public, Protective, and Government Services
  - 16: Hospitality and Tourism
  - 17: Architecture and Construction
  - 18: Agriculture, Food, and Natural Resources
  - 19: Human Services
  - 20: Transportation, Distribution, and Logistics
  - 21: Engineering and Technology
  - 22: Miscellaneous
  - 23: Non-Subject Specific [NOT USED FOR HSTS]
  - 24: World Languages

Examples of High School Course Catalogs

Grade 9	Grade 10	Grade 11	Grade 12
Ancient Literature	Early Modern Literature	American Literature (AP English Language Composition)	Modern European Literature (AP English Literature & Composition)
Ancient History	Medieval/Early Modern History	American History	Modern European History
Composition	Logic & Rhetoric	American Government, Economics	Moral Philosophy & Senior Tutorial
Algebra I or Geometry	Geometry or Algebra II	Algebra II or Pre-Calculus	Pre-Calculus or AP Calculus
Biology	Chemistry or Integrated Physics/Chemistry	Physics or Astronomy, or AP Science (Biology, Chemistry, Physics)	Physics or Astronomy, or AP Science (Biology, Chemistry, Physics)
Latin and/or French	Latin, French or Greek	Latin, French or Greek	Latin, French or Greek
Physical Education, Health, Fine Arts + additional electives			



High School Course Guide -

Acting Basics

SCHOOLS:	Roosevelt
SUBJECT AREA:	Visual and Performing Arts
a-g DESIGNATION:	f - Visual and Performing Arts (VAPA)
GRADES:	10, 11, 12
PREREQUISITE:	Introduction to Theatre or instructor permission

A course geared to the intermediate actor, centering around theatre games, exercises, audition techniques, improvisation, monologue, scene, and one-act work.

Acting Styles

SCHOOLS:	Roosevelt
SUBJECT AREA:	Visual and Performing Arts
a-g DESIGNATION:	f - Visual and Performing Arts (VAPA)
GRADES:	10, 11, 12
PREREQUISITE:	Acting Basics or instructor recommendation

A continuation of Acting Basics together with advanced interpretation, criticism, and acting based on a historical framework. Students learn and apply acting techniques from all the major periods of theatre.

Algebra / Geometry III

SCHOOLS:	Bullard, Cambridge, Design Science, De Wolf, Duncan, Edison, Fresno, Hoover, J.E. Young, McLane, Patino, Roosevelt, Sunnyside
SUBJECT AREA:	Mathematics
a-g DESIGNATION:	c - Mathematics
GRADES:	11, 12
PREREQUISITE:	Algebra I and Geometry

This third year integrated math course includes major topics such as: operations with whole numbers, solving equations, including quadratic equations, geometric reasoning, similarities and congruencies, probability, statistics, transformations, and trigonometry.

Algebra I (CCSS)

SCHOOLS:	Bullard, Cambridge, Design Science, De Wolf, Duncan, Edison, Fresno, Hoover, J.E. Young, McLane, Phoenix, Roosevelt, Sunnyside
SUBJECT AREA:	Mathematics
a-g DESIGNATION:	c - Mathematics
GRADES:	9, 10
PREREQUISITE:	None

Algebra I is the foundation course for all higher mathematics courses and emphasizes the learning of essential concepts which are required for further success in mathematics. Topics include: operations with integers, solving equations and inequalities, exponents, operations with polynomials, graphing in two variables, systems of equations, rational algebraic expressions, and application problems.

Algebra I SDC

SCHOOLS:	Bullard, Edison, Fresno, Hoover, McLane, Roosevelt, Sunnyside
SUBJECT AREA:	Mathematics
a-g DESIGNATION:	Non a-g
GRADES:	9, 10, 11, 12
PREREQUISITE:	None

hscg.fresnounified.org Page 1 of 39

Course Catalog Coding in Large-Scale Survey Operations

- Convert locally defined course information into standardized SCED codes.

What is being coded?

- Course titles
- Course descriptions
- Credit and sequence information
- Grade level
- Flags for special education, course language taught, and online courses.

Why is this step important?

- Supports consistent classification of courses across education agencies.
- Supports national reporting and analysis.
- Central to data quality, consistency, and interpretability.

Operational Challenges in Qualitative Coding

- **Variation in Course Information**
  - Inconsistent course titles across education agencies.
  - Limited or uneven course descriptions.
- **Reliance on Human Judgment**
  - Interpretation required to determine course content.
  - Potential for inconsistency across coders.
- **Scale of Processing**
  - Hundreds of course catalogs.
  - Resulting in tens of thousands of course records to code per cycle.
- **Operational Burden**
  - Time-intensive manual coding.
  - Training and quality control requirements.

Approximate volume of data collected by HSTS in 2019

Count	2019
High schools	~1,400
High school graduates	~47,300
↓	↓
Catalogs	~900
Catalog courses	~344,000
Transcript courses	~2.3 million

Sources: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP) High School Transcript Study (HSTS), 2009 and 2019.

Challenging Cases Requiring Human Judgment

Ambiguous course titles

E.g., Advanced Math Topics, Integrated Science

Locally defined or non-standard courses

E.g., STEM Seminar, College Readiness Mathematics

Multi-subject or interdisciplinary courses

E.g., Physics of Music, Environmental Science & Policy

Courses with insufficient or vague descriptions

Limited detail in catalogs or transcripts.

New or evolving course offerings

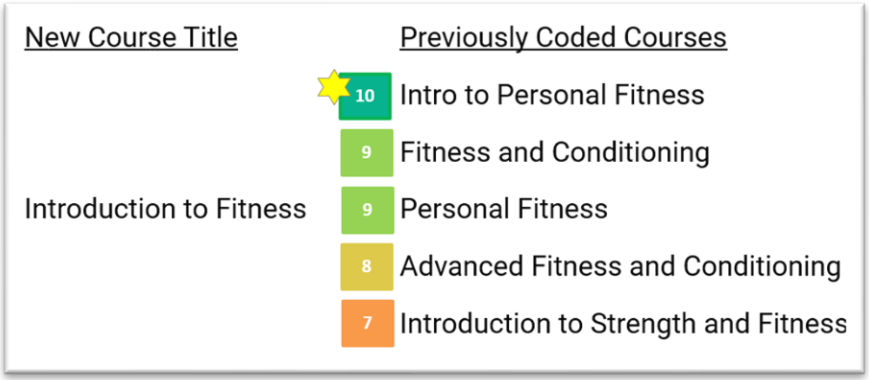
Courses are not well represented in the SCED coding system yet.

Applying Machine Learning (ML) Models to Support Standardized Coding

- Represent course titles and descriptions as semantic embeddings using natural language processing (NLP), which uses ML techniques.
- Capture contextual meaning beyond keyword matching.
- Compare each course to a database of previously coded HSTS courses.
- Compute similarity scores to identify likely SCED codes.
- Provide ranked suggestions (top-k) to support human coding decisions.
- More robust than lexical or rule-based approaches.
- Human coders retain full discretion, including selecting codes outside model suggestions.

Applying NLP in HSTS

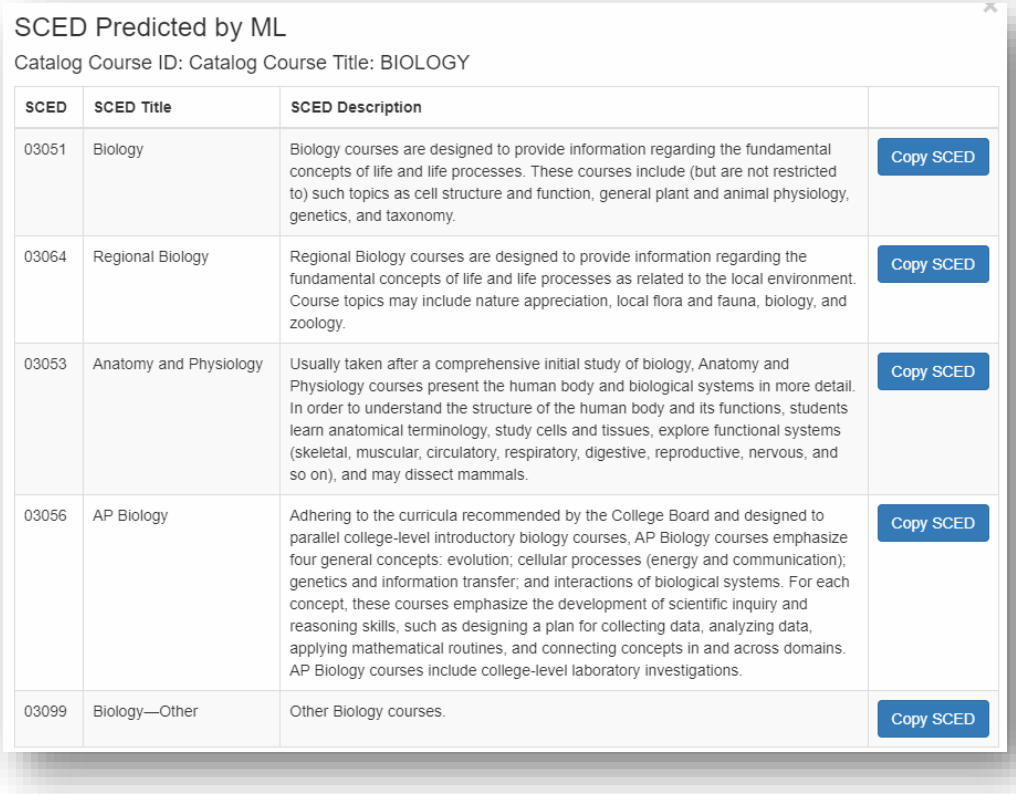
- Purpose: SCED code output for course title input.
- Dataset: HSTS 2019 coded catalog courses.
- Generates embeddings for all titles in the 2019 dataset.
- Generate and compare input embedding to 2019 dataset embeddings to find the most similar ones (similarity scores are generated for 2019 course titles).



Course Title	State	SCED Code
Algebra 1	Texas	02052
Introduction to Fitness	Georgia	08016
Principles of Engineering	Kansas	21004
Advanced Composition	Michigan	01103
Spanish III	Maryland	24056

Applying NLP in HSTS (cont.)

- Potential 2019 course title matches are linked to their SCED codes.
- Coders are presented with potential SCED codes.
- For initial testing, 5 matches were shown.
  - Precedent from other NCES projects using matching programs.



SCED Predicted by ML

Catalog Course ID: Catalog Course Title: BIOLOGY

SCED	SCED Title	SCED Description	
03051	Biology	Biology courses are designed to provide information regarding the fundamental concepts of life and life processes. These courses include (but are not restricted to) such topics as cell structure and function, general plant and animal physiology, genetics, and taxonomy.	Copy SCED
03064	Regional Biology	Regional Biology courses are designed to provide information regarding the fundamental concepts of life and life processes as related to the local environment. Course topics may include nature appreciation, local flora and fauna, biology, and zoology.	Copy SCED
03053	Anatomy and Physiology	Usually taken after a comprehensive initial study of biology, Anatomy and Physiology courses present the human body and biological systems in more detail. In order to understand the structure of the human body and its functions, students learn anatomical terminology, study cells and tissues, explore functional systems (skeletal, muscular, circulatory, respiratory, digestive, reproductive, nervous, and so on), and may dissect mammals.	Copy SCED
03056	AP Biology	Adhering to the curricula recommended by the College Board and designed to parallel college-level introductory biology courses, AP Biology courses emphasize four general concepts: evolution; cellular processes (energy and communication); genetics and information transfer; and interactions of biological systems. For each concept, these courses emphasize the development of scientific inquiry and reasoning skills, such as designing a plan for collecting data, analyzing data, applying mathematical routines, and connecting concepts in and across domains. AP Biology courses include college-level laboratory investigations.	Copy SCED
03099	Biology—Other	Other Biology courses.	Copy SCED

Example of ML-Supported Course Coding – AP BIOLOGY

SCED Predicted by ML

Catalog Course ID: Catalog Course Title: AP BIOLOGY

SCED	SCED Title	SCED Description	
03056	AP Biology	Adhering to the curricula recommended by the College Board and designed to parallel college-level introductory biology courses, AP Biology courses emphasize four general concepts: evolution; cellular processes (energy and communication); genetics and information transfer; and interactions of biological systems. For each concept, these courses emphasize the development of scientific inquiry and reasoning skills, such as designing a plan for collecting data, analyzing data, applying mathematical routines, and connecting concepts in and across domains. AP Biology courses include college-level laboratory investigations.	Copy SCED
03051	Biology	Biology courses are designed to provide information regarding the fundamental concepts of life and life processes. These courses include (but are not restricted to) such topics as cell structure and function, general plant and animal physiology, genetics, and taxonomy.	Copy SCED
03052	Biology-Advanced Studies	Usually taken after a comprehensive initial study of biology, Biology-Advanced Studies courses cover biological systems in more detail. Topics that may be explored include cell organization, function, and reproduction; energy transformation; human anatomy and physiology; and the evolution and adaptation of organisms.	Copy SCED
03999	Life and Physical Sciences-Other	Other Life and Physical Sciences courses.	Copy SCED
03064	Regional Biology	Regional Biology courses are designed to provide information regarding the fundamental concepts of life and life processes as related to the local environment. Course topics may include nature appreciation, local flora and fauna, biology, and zoology.	Copy SCED

Multi-Phase Experimental Design of ML approach in HSTS

System Integration

- Integrated ML into the computing environment and coding database.
- Evaluated compatibility with existing workflow.

Output Validation

- Compared model suggestions to final SCED codes.
  - **72%** first suggestions matched the final codes.
  - **91%** of the top five included correct codes.
- Coders could select codes outside the model's suggestions.

User Testing

- Experienced coders coded with and without ML support.
- Assessed impact on accuracy and efficiency.

User Testing

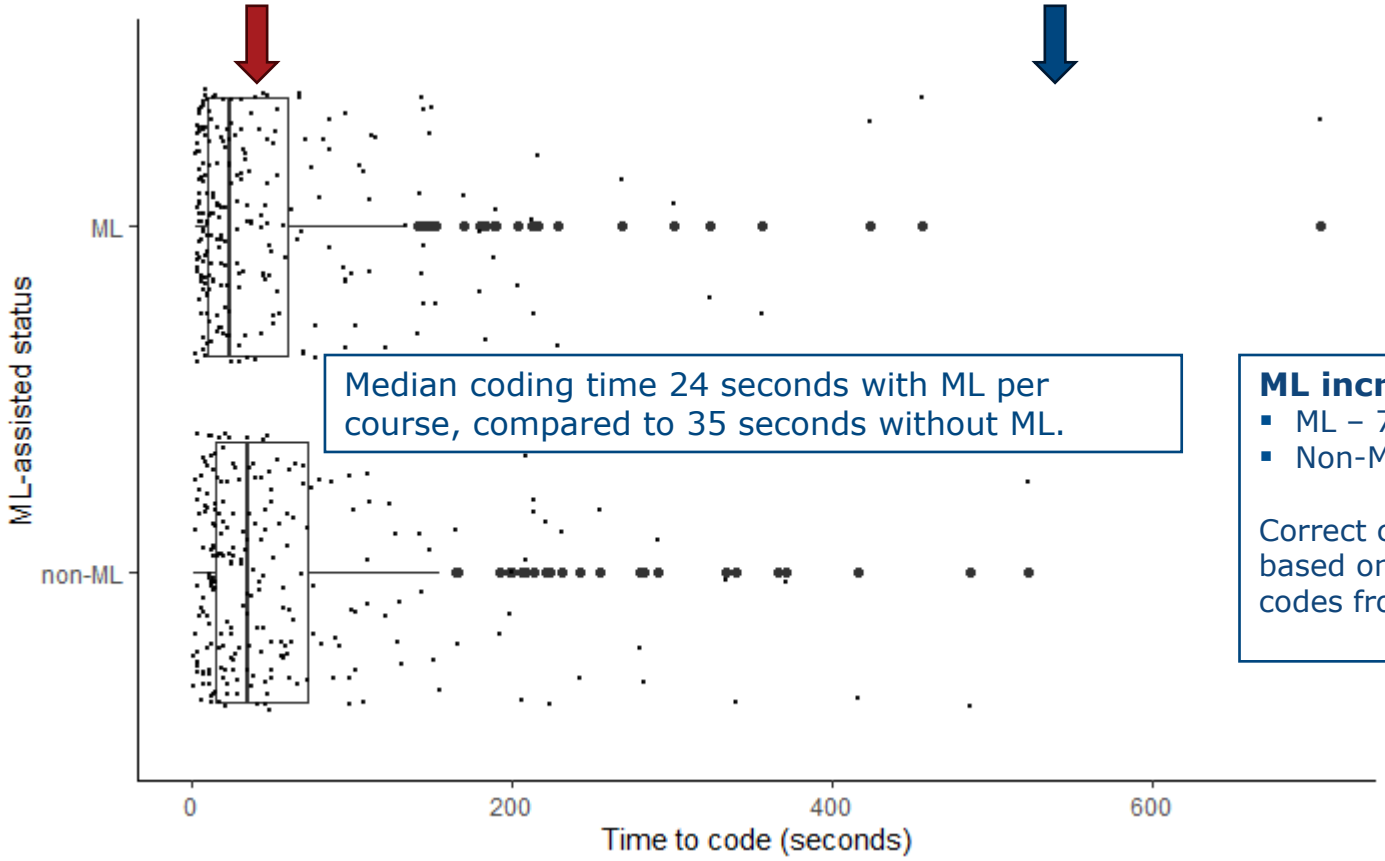
Counterbalancing experimental design with 2 treatment groups

Course Group	ML	Not ML
1	Coder A	Coder B
2	Coder C	Coder D
3	Coder B	Coder A
4	Coder D	Coder C
5	Coder A	Coder C
6	Coder B	Coder D
7	Coder C	Coder A
8	Coder D	Coder B
9	Coder A	Coder C
10	Coder B	Coder D
11	Coder C	Coder A
12	Coder D	Coder B

Analyses focus on questions:

- Does coding time improve using ML?
  - Mean coding time per course.
- Does coding accuracy improve using ML?
  - Proportion of courses with codes in agreement with the correct answers.
- Do the users feel comfortable using ML?

User Testing – Results on Coding Speed and Accuracy



Median coding time 24 seconds with ML per course, compared to 35 seconds without ML.

ML increased accuracy:

- ML – 74% correct codes.
- Non-ML – 66% correct codes.

Correct codes were defined based on the assigned SCED codes from the HSTS 2019 cycle.

Lessons Learned from ML-Supported Course Coding

- Human + ML performs better than either alone.
- The top-k suggestions are more effective than single predictions.
- Performance gains are strongest for routine, high-frequency courses.
- Human oversight remains essential for ambiguous or novel courses.
- System performance depends on the quality and coverage of training data.

Implications for Data Quality in Federal Surveys

- Improves coding **accuracy, consistency, and efficiency**, supporting faster processing and more cost-effective operations.
- Reduces **variability** in the **interpretation** of course descriptions.
- Enhances the **reproducibility** of coding decisions.
- Supports **standardization** across course codes.
- Aligns with federal data quality standards, emphasizing **reliability**, **comparability**, and **transparency**.

Thank you!

Judy H. Tang

JudyTang@Westat.com