Psychometric Challenges and Opportunities in the Evolution of STAR

Robert Anderson, California Department of Education

Denny Way, Educational Testing Service

Presentation at the California Educational Research Association
81st Annual Meeting, November 14, 2002

Overview of Presentation

- Two Presenters (and perspectives):
  - A technical presenter: Denny Way, ETS Director of Psychometrics for STAR and CAHSEE programs
  - Discussant from a content and policy perspective: Bob Anderson, California Department of Education

Overview of Presentation

- Three Presentation Topics
  - Standards-based Test Development: What is the best way to measure rigorous content standards given students at a variety of proficiency levels?
  - Vertical Scaling the CSTs: What are the pros and cons of vertical scales for the CSTs?
  - Linking the Stanford/9 and CAT/6 norm-referenced tests: What are the technical plans for the 2003 STAR?

Overview of the Presentation

- Structure of the presentation
  - Technical information and analyses first
  - Content and policy perspectives next
  - Questions, comments, discussion from the audience

Topic 1: Balancing Content Rigor and Psychometrics in Standards-Based Assessment

- The CSTs are aligned with state-adopted standards that describe what California students should know and be able to do in each grade and content area tested
- The CSTs report student performance with respect to five performance levels
  – Advanced, Proficient, Basic, Below Basic, and Far Below Basic

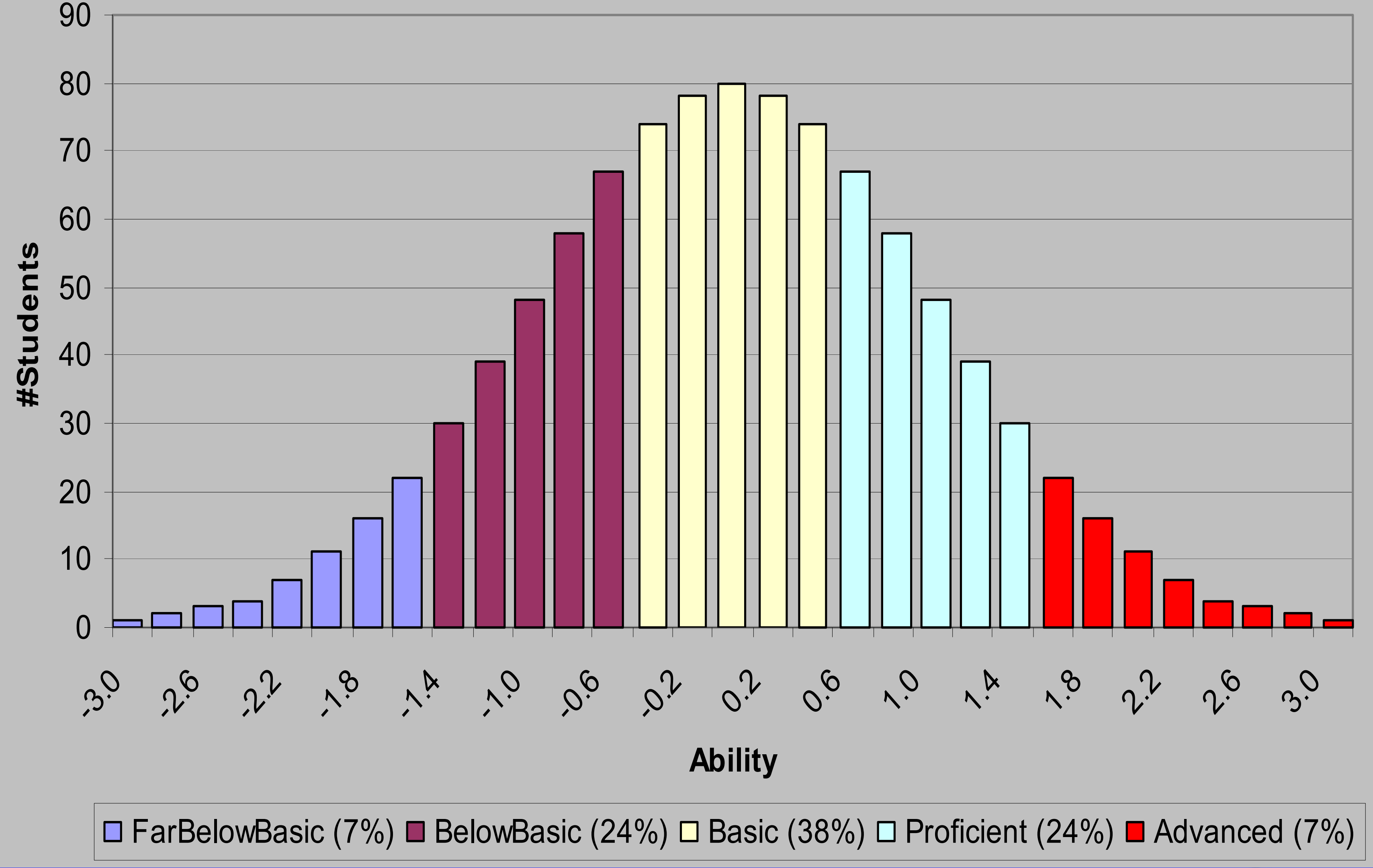
Topic 1: Balancing Content Rigor and Psychometrics in Standards-Based Assessment (Continued)

- Review of the standards suggests that questions must be sufficiently challenging to appropriately measure the standards
- Question difficulty can be modified, and questions of varying difficulty can each provide valid measurement of the same content standard

Analyzing Test Difficulty and Psychometric Characteristics: An Illustrative Study

- "What if" analysis of differences in item difficulty
- Hypothesized a normally distributed student population with "simulated" students at five levels of proficiency
- Used Item Response Theory (IRT) to simulate different test characteristics
  - Level of average item difficulty
  - Variability in item difficulty
- Evaluated the results in terms of errors of measurement and reliability

Distribution of Student Ability



Characteristics of Three Possible Tests (All Tests with 75 Items)

- A difficult test with few extremely easy or difficult items
- An easier test with a normal range of item difficulties
- An easier test with a wide range of item difficulties

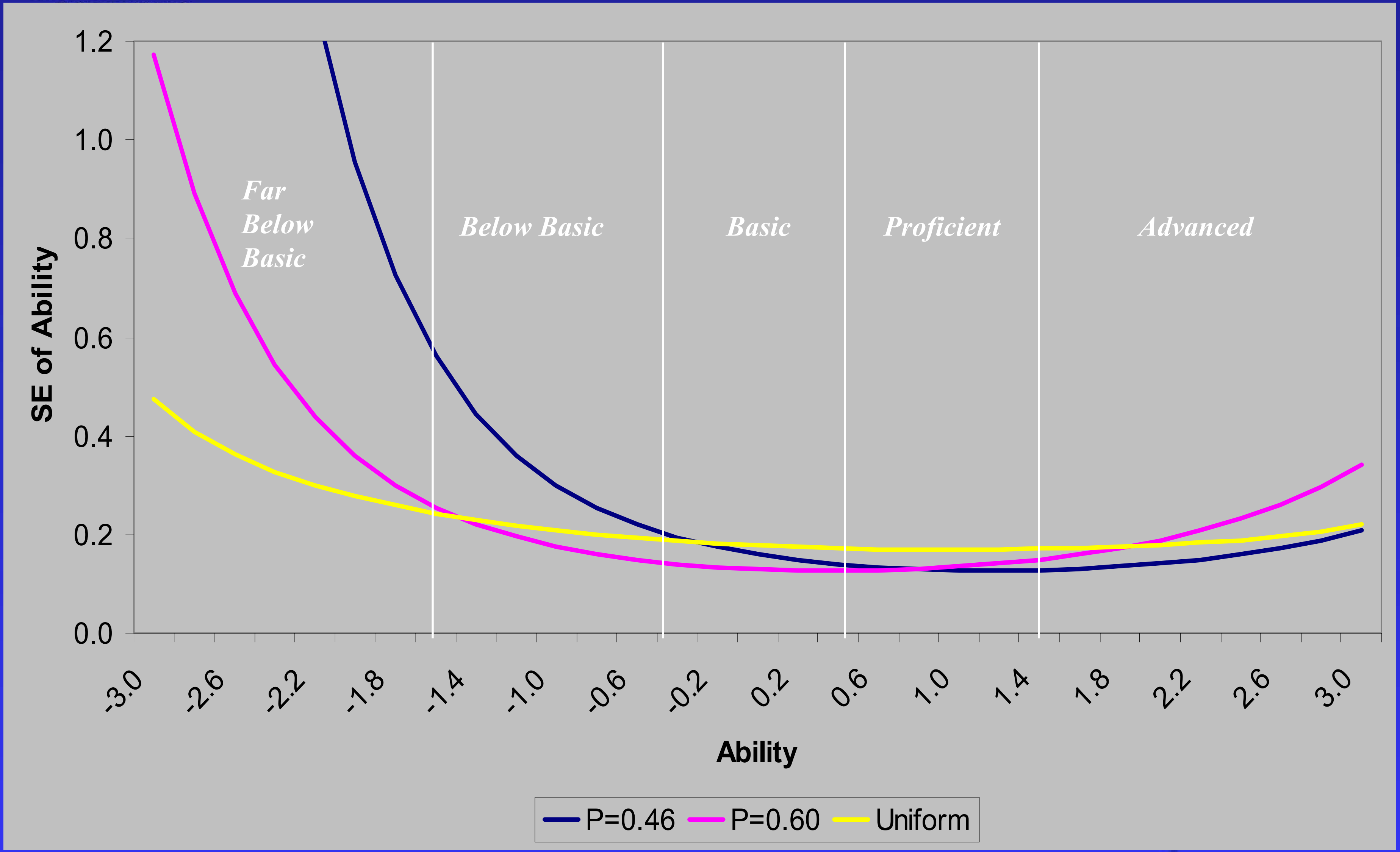
Characteristics of Three Possible Tests

	Difficult	Normal	Uniform
Mean P-Value	0.46	0.60	0.60
Std P-Value	0.16	0.18	0.26
Min P-Value	0.22	0.25	0.22
Max P-Value	0.80	0.95	0.98

Some Technical Details

- Used 1PL Model with constant guessing parameter of 0.20
- Calculated “score information function” at 31 ability levels from –3.0 to 3.0 by 0.2
- Conditional standard errors equal to the square root of the reciprocal of the score information function
- Simulated 1000 cases and estimated test characteristics (mean, SD, reliability)

Conditional Standard Errors of Three Possible Tests



Technical Characteristics of three Possible Tests

	Difficult	Normal	Uniform
Number Items	75	75	75
Test Mean:	33.56	44.64	44.69
Test SD	13.47	14.38	10.83
Reliability:	0.92	0.94	0.90

Educational Testing Service

Topic 1 Conclusions

- Different psychometric characteristics of tests result in errors of measurement at different points of the score scale
- Standards-based tests are challenging to create because it is difficult to measure well at every proficiency level
- Technical test specifications should be set with knowledge of the psychometric implications

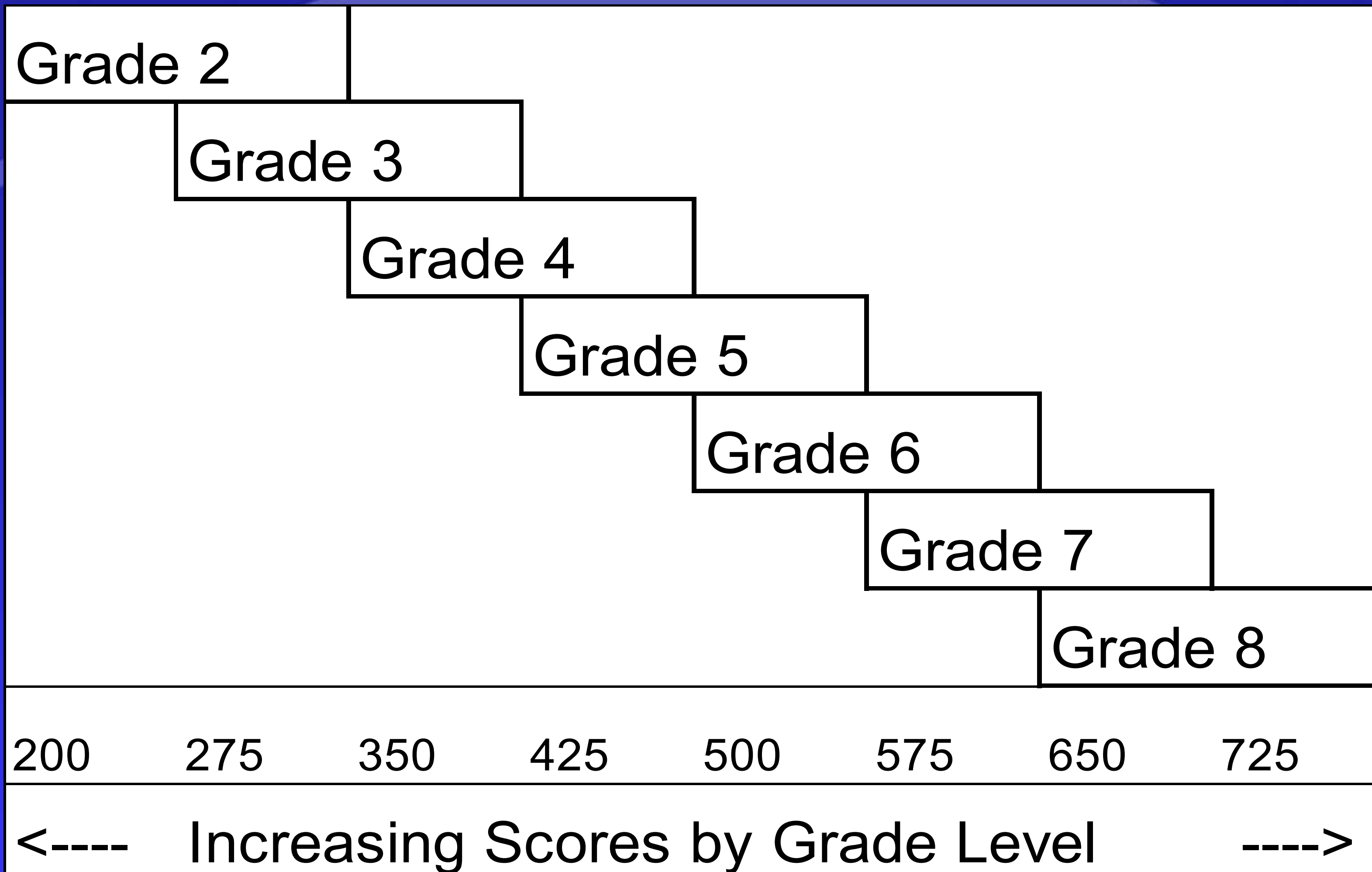
Topic 2: Developing Vertical Scales for the CSTs

- Current CSTs have unique scales for each grade and content area
- This means that scores may only be compared for the same grade and content area
- Norm-referenced tests provide vertical scales for comparing student growth, but NRTs do not measure the state content standards

What is a Vertical Scale?

- Also referred to as a “developmental scale”
- Content defines a “developmental continuum” for a particular area
- With developmental scales, tests used in different grades are calibrated against one another
- Developmental scales facilitate the estimation of individual growth and the use of individualized test administration

Example of a Vertical Scale



Issues with Vertical Scales

- Vertical scales are only sensible when the content supports interpretations across levels
- Scales suggest interpretations that may not be appropriate
- Vertical scaling differs from test equating in that the linking does not assume parallel forms
  - Scaling methodology an important consideration
  - Content dimensionality may have a greater impact

Vertical Scales and Content Considerations

- Standards more closely aligned across grades:
  - English Language Arts
  - Mathematics from Grader 2 to 7
- Standards less closely aligned across grades:
  - End of Course Mathematics Tests
  - End of Course Science Tests
  - End of Course History and Social Science Tests

What Has to Happen to Produce Vertical Scales for the CSTs?

- Policy decision must be made
- CSTs must include common items across adjacent grades
- New score scales must be defined and technical work to accomplish vertical scaling must be done
- Issues related to within-grade proficiency levels must be addressed
- Could be done as part of 2004 STAR

Topic 3: Linking the Stanford / 9 and CAT / 6 Scales

- Background
- ETS Technical Plans
  - Analyses prior to 2003 STAR administration
  - Analyses as part of the 2003 STAR administration

SAT/9 – CAT/6 Linking Study: Background

- The Academic Performance Index (API) includes contributions from NRT results
- NRT contributions have decreased over the past three years
- The API has had increasing contributions from the CSTs and now includes CAHSEE results as well

Evolution of the API

- Base 1999 (Growth 2000)
  - 100% Stanford 9
- Base 2000 (Growth 2001)
  - 100% Stanford 9
- Base 2001 (Growth 2002)
  - 64% Stanford 9
  - 36% CSTs
- Base 2002 (Growth 2003)
  - 29-40% Stanford 9
  - 56-60% CSTs
  - 0-15% CAHSEE
- Beyond 2003
  - CST Science and History/Social Science
  - CAPA
  - Attendance Rates
  - Graduation Rates

Linking NRTs for the API

- Any new NRT chosen for 2003 would have required a linking study
- ETS proposal included use of the California Achievement Test, Version 6 (CAT/6)
- Linking study is needed to provide concordance between CAT/6 scores and Stanford 9 equivalents in order to calculate the NRT contribution to API growth

Linking Study Designs

- Two linking study options
  - Direct study where SAT/9 and CAT/6 are administered "together"
  - Indirect study where SAT/9 and and CAT/6 scores are linked using CST test scores as an external anchor
- Indirect study chosen because of cost, feasibility, and technical considerations

ETS Technical Linking Plans

- Linking will be done using equipercentile methodology
- CSTs will be used as a common anchor between SAT/9 scores obtained in 2002 and CAT/6 scores obtained in 2003
- Study will have two parts
  - Initial investigations of SAT/9 and CSTs
  - Complete linkings once CAT/6 scores are available

Some Technical Details

- Use of the CSTs in the linking
  - Clearly defined for ELA and grades 2-7 mathematics
  - More complicated for high school math and science tests
- Analyses of subpopulations
  - Provides detail about invariance of the overall linkings
  - Comparisons can include specified subgroups, school districts, and even individual schools

Some Correlation Data for High School Mathematics

CST Content Area	Correlation with SAT/9
First Year Integrated Math	0.63
Algebra I	0.64
Second Year Integrated Math	0.76
Geometry	0.77
Third Year Integrated Math	0.72
Algebra II	0.75
High School Summative Math	0.85

Closing Notes on the Linking Study

- ETS technical gurus will provide consultation on linking study analyses
- Initial analyses will be reviewed by CDE and their API advisors
- Final analyses will be completed over the summer after sufficient data from STAR 2003 are available

Further Technical Information

Denny Way

dway@ets.org

Complaints, criticisms, discussion of any unpleasant topics

George Powell

gpowell@ets.org