LANGUAGE TESTING FORUM (LTF)

University of Bedfordshire, 24 November 2018

A Bayesian approach to improving measurement precision over multiple test occasions

ALISTAIR VAN MOERE & SEAN HANLON

MetaMetrics

LEXILE

MetaMetrics

Single-administration Tests

MetaMetrics.

In educational measurement, there is a tendency to interpret scores from a single administration of a test as an accurate indicator of an underlying latent trait.

Usually, test scores are taken at face value, even though scores fluctuate over test occasions, e.g.

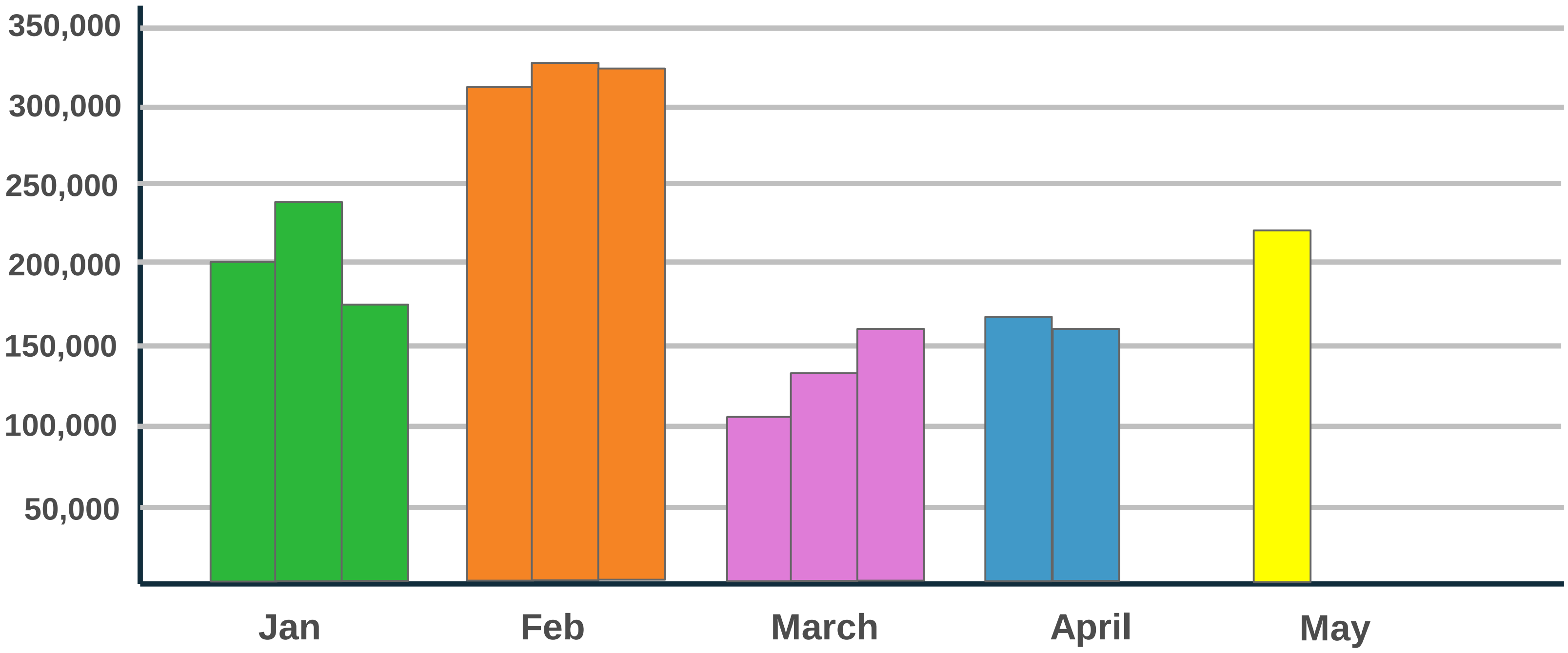
- immigration or university entrance tests
- formative assessments administered multiple-times-per-year

However, other fields make more use of prior data, or make adjustments to their estimates as new data comes to light

LEXILE®

Non-Farm Payroll

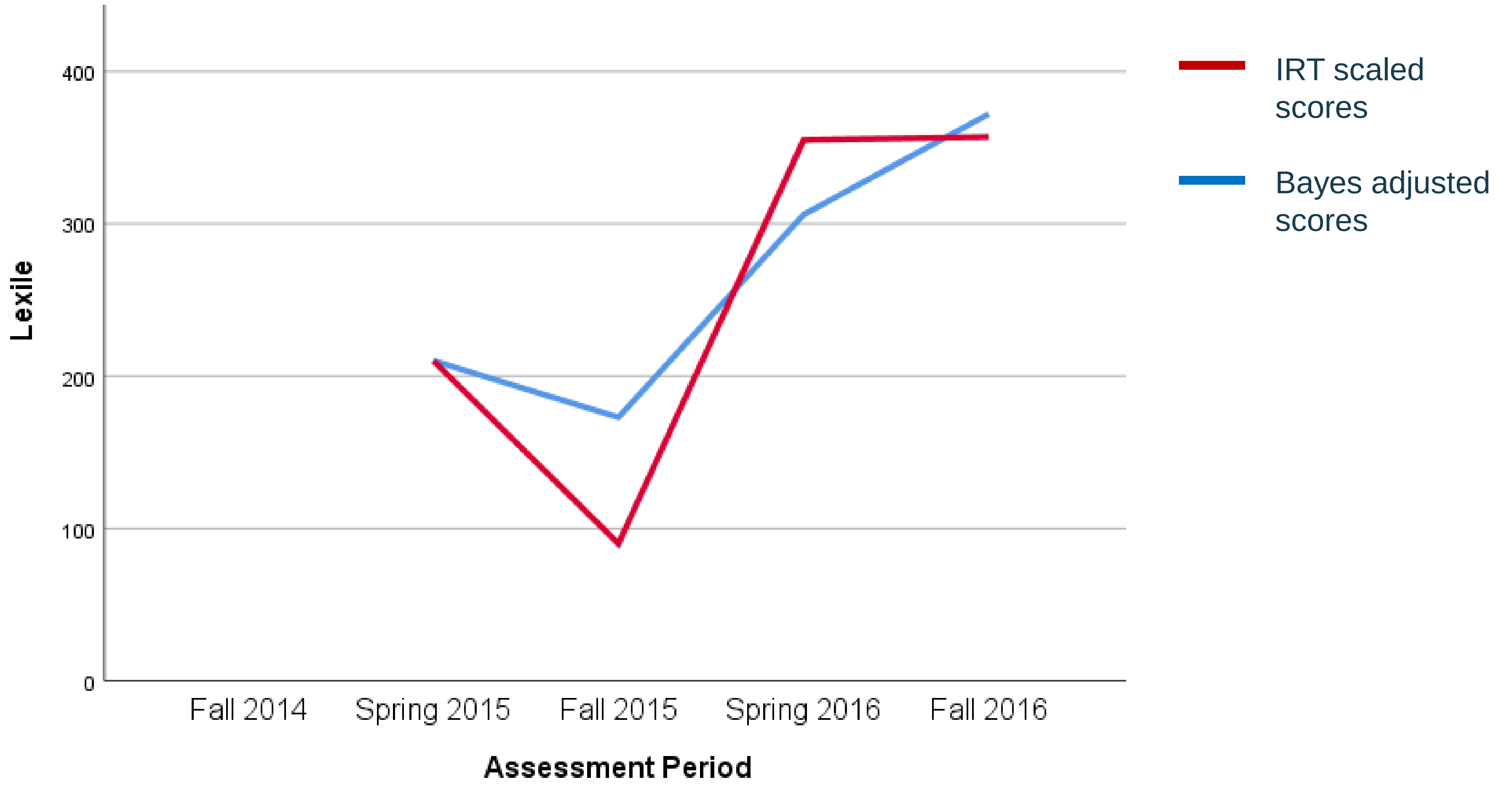
When the US Bureau of Labor Statistics reports this data monthly, they update the previous month, and the month before that.



Combining recent and prior scores

This paper researches a statistical approach, using Bayes theorem, to combine previous test scores with new test scores in order to arrive at a more-precise estimate of student ability.

Combining recent and prior scores



Two Approaches to estimating student ability

MetaMetrics.

Familiar approach

Raw score → **IRT modeling** → **IRT Scaled score with standard error**

Bayes approach

IRT Scaled score + uncertainty → **IRT Scaled score + uncertainty** → **Ability estimate + uncertainty**

Prior ability estimate & recent estimate ability = new ability estimate

LEXILE®

Steps for computation

1. Get previous values:

   Student comes into the testing session with b_{prior} and σ_{prior} from t_0.

2. Update values:

   At time $t_{1,}$ student takes the test and b_{update} and σ_{update} are computed.

   b_{update} takes account of learning gains since the prior estimate
   σ_{update} takes account of increasing uncertainty since the prior estimate

3. Compute new values:

   Combine updated prior information with current information.

Customizable features in the model

MetaMetrics.

1. Incorporating growth or learning gains

An assumption can be made that test-takers' skills are improving over time

2. Increasing uncertainty over time

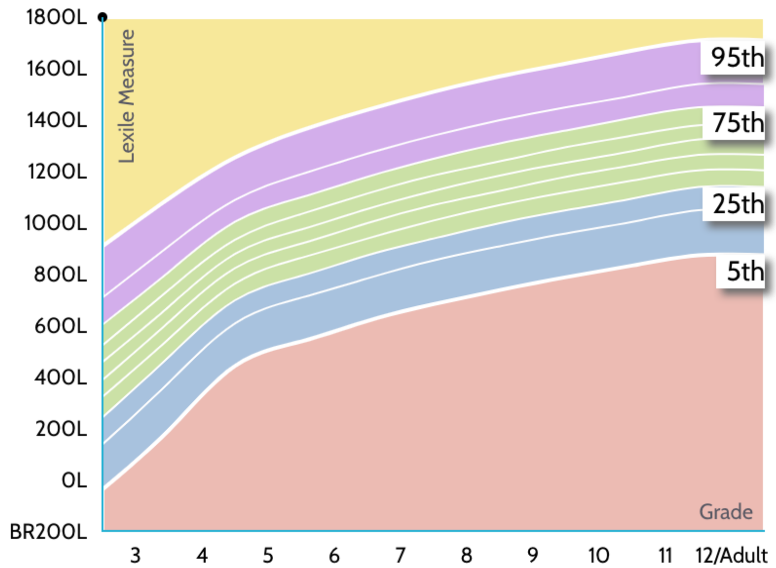
An assumption can be made that *the more time passes* between tests, the *less useful* prior scores are

Customizable features

1. Growth assumption

When average growth rates have been estimated for a known population, these can be incorporated into the model.

Typical Rates of Reading Growth in US Schools



Customizable features

MetaMetrics.

1. Growth assumption

When average growth rates have been estimated for a known population, these can be incorporated into the model.

Student grade or year	Average measures	Average annual progress
3rd to 4th	664L to 803L	139L
4th to 5th	803L to 925L	122L
5th to 6th	925L to 1029L	104L

Williamson, G. (2016). Novel Interpretations of Academic Growth. JAEPR, 2(2), 15-35. ISSN: 1930-9325

LEXILE®

Customizable features

Growth assumption

Younger, low ability students tend to grow faster than older, more experienced students. Variable rates of growth can be modeled for specific populations.

$$b_{update} = b_{prior} + g_s(t_1 - t_0)$$

Based on a known growth curve

Customizable features

MetaMetrics.

2. Increase of uncertainty over time

- Uncertainty from a prior test increases as time passes (i.e. uncertainty from a test taken 12 months ago is higher than a test taken 6 months ago.)
- After how many months should we say that the information in a prior test has no value? (i.e. at what time period do we set maximum uncertainty?)

For example:

$$\sigma_{update} = \frac{(\sigma_{maximum} - \sigma_{prior})(t_2 - t_1)}{t_{maximum}} + \sigma_{prior}$$

Customizable features

MetaMetrics.

2. Increase of uncertainty over time

- Uncertainty from a prior test increases as time passes (i.e. uncertainty from a test taken 12 months ago is higher than a test taken 6 months ago.)
- After how many months should we say that the information in a prior test has no value? (i.e. at what time period do we set maximum uncertainty?)

For example:

$$\sigma_{update} = \frac{(\sigma_{maximum} - \sigma_{prior})(t_2 - t_1)}{1095.75} + \sigma_{prior}$$

$\uparrow$

Number of days in 3 years

LEXILE®

Research question

To what extent do students' reading ability estimates differ when using Rasch-scaled scores from a single administration *versus* Bayesian scoring with priors?

Contexts:

- L2 students in an after-school English-language instructional program in S Korea (n=25)
- L1 elementary students in schools in the US following a curriculum-based reading program (n=20,928)

Assessment instruments

Naturalistic data was obtained from two the assessment contexts

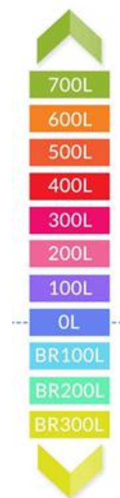
- Progress monitoring tests designed to measure reading comprehension
- Fixed-form, 30-40 multiple choice items in each test
- Specific test forms developed for different levels, as aligned with instructional program
- Scores reported on the Lexile Scale

Lexile Score Scale

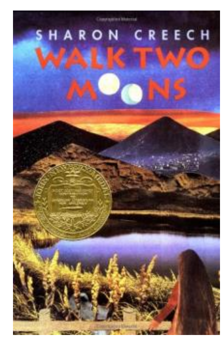
Lexile measures:

- A framework for connecting readers with level-appropriate texts
- An equal interval scale, from below zero to above 2,000
- **Lexile *student* measures** refer to reader ability
- **Lexile *text* measures** refer to text complexity

770L reader



770L book



Lexile Score Scale

- Over 65 reading assessments have been linked to this scale, e.g. by seeding items from an anchor item bank into their assessments
- 100 million books, articles and websites have been measured
- Approx 35 million students receive Lexile measures globally

White, S., & Clement, J. (2001). *Assessing the Lexile Framework: Results of a panel meeting (NCES 2001-08)*. Washington, DC: U.S. Department of Education, Office of Educational Research and Improvement.

Lexile Score Scale

- Over 65 reading assessments have been linked to this scale, e.g. by seeding items from an anchor item bank into their assessments
- 100 million books, articles and websites have been measured
- Approx 35 million students receive Lexile measures globally

Scientists have made a discovery about the moon. They believe its interior contains much more water than previously thought, though exactly how much is still unclear. If scientists are correct, future astronauts may benefit from this finding. Traveling through the solar system requires extensive supplies, including water. With improved technology, astronauts could extract water from the moon and leave water from Earth off their packing lists.	The discovery could be: ○ misleading ○ advantageous ○ symbolic ○ costly

Research question

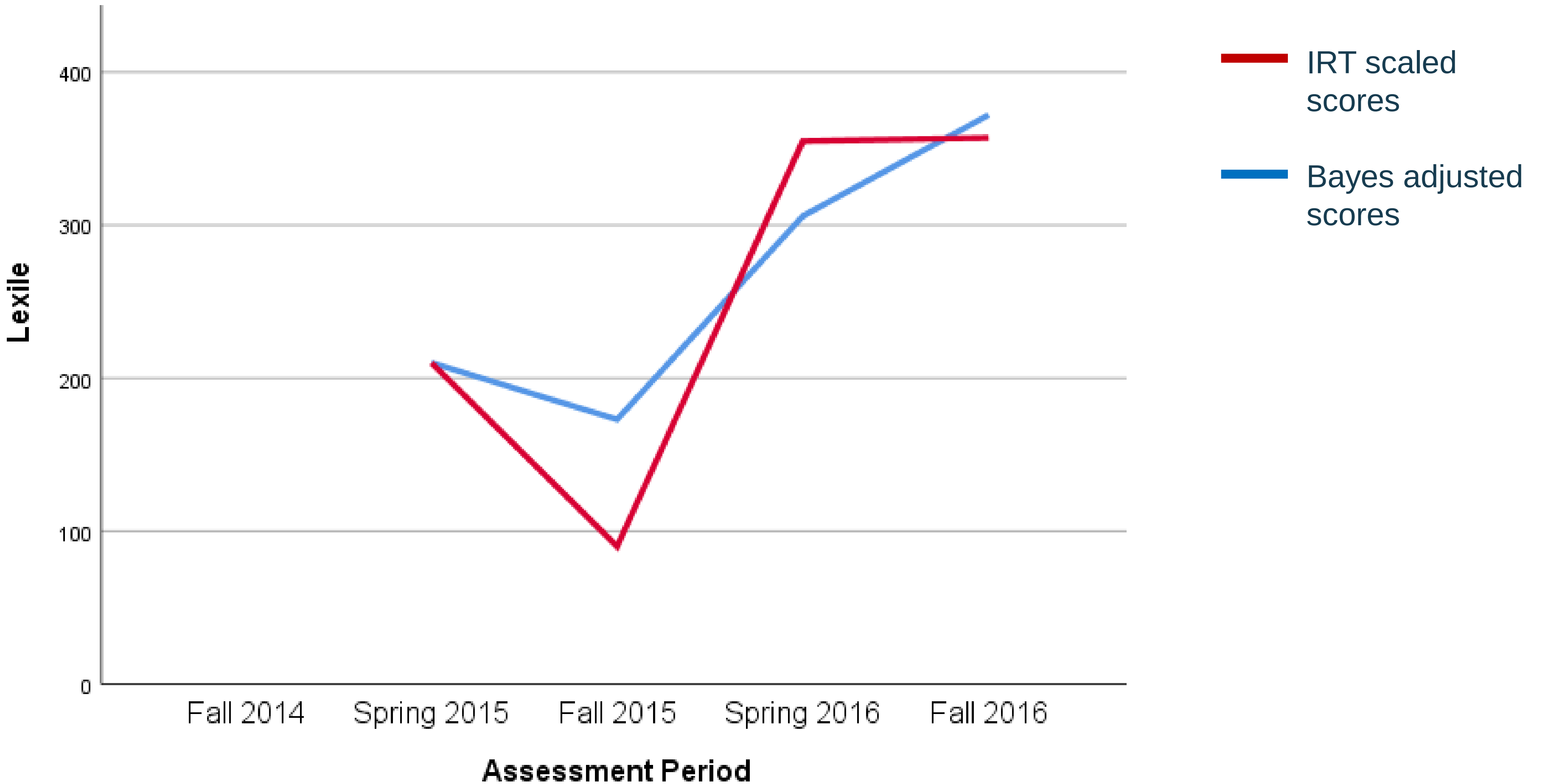
MetaMetrics.

To what extent do students' reading ability estimates differ when using Rasch-scaled scores from a single administration *versus* Bayesian scoring with priors?

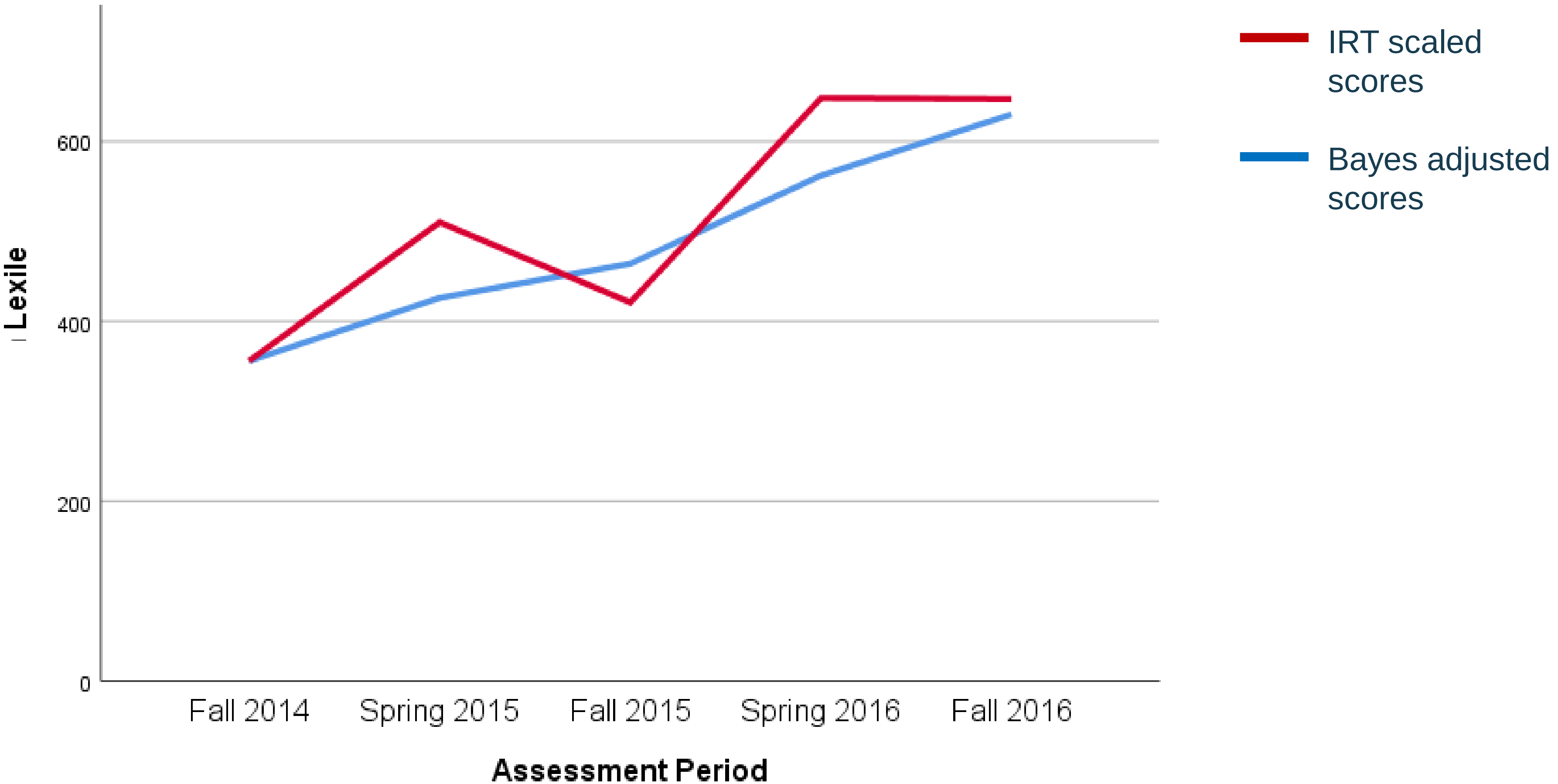
Contexts:

- L2 students in an after-school English-language instructional program in S Korea (n=25)
- L1 elementary students in schools in the US following a curriculum-based reading program (n=20,928)

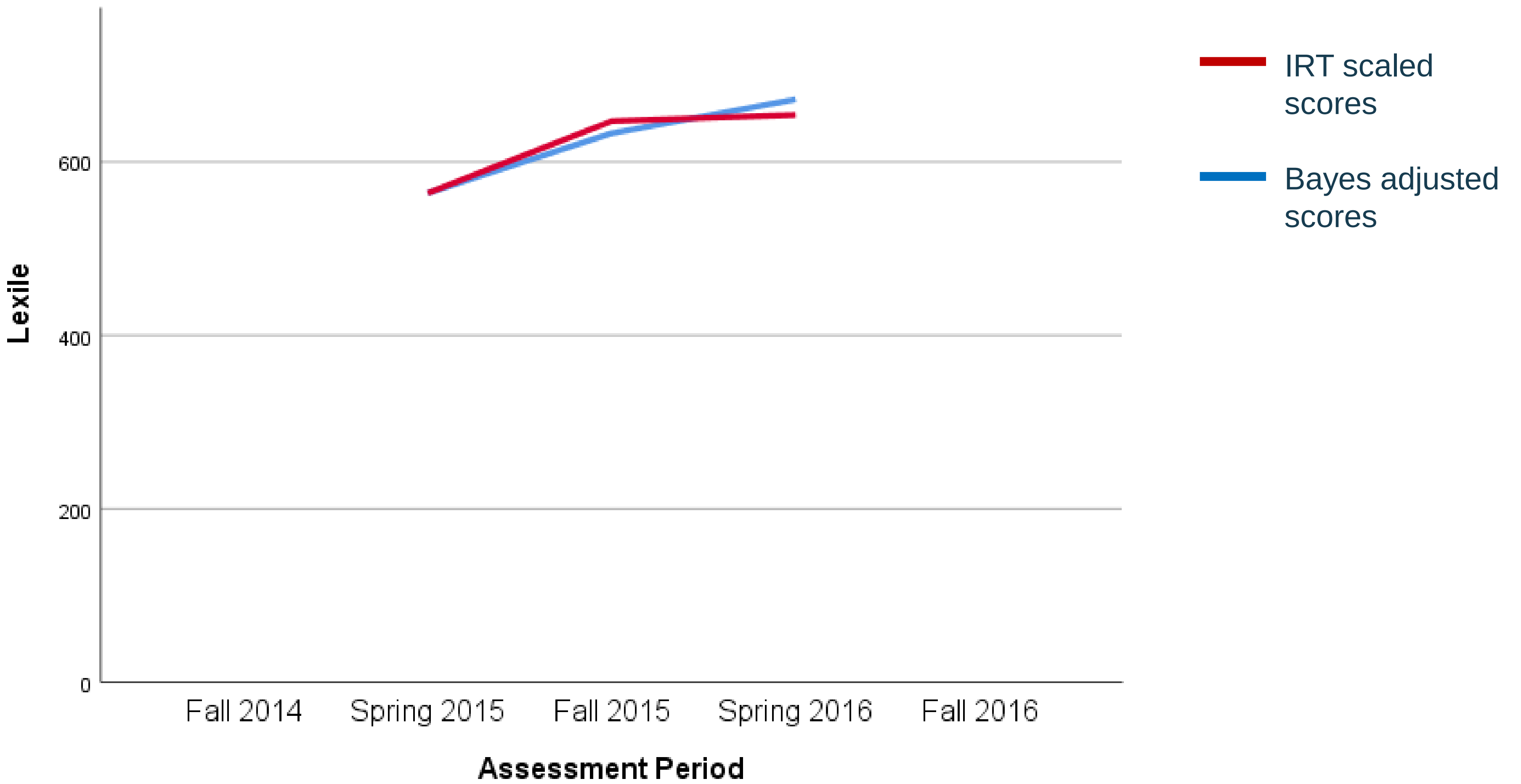
Example: Bayes "smooths out" a student's measures



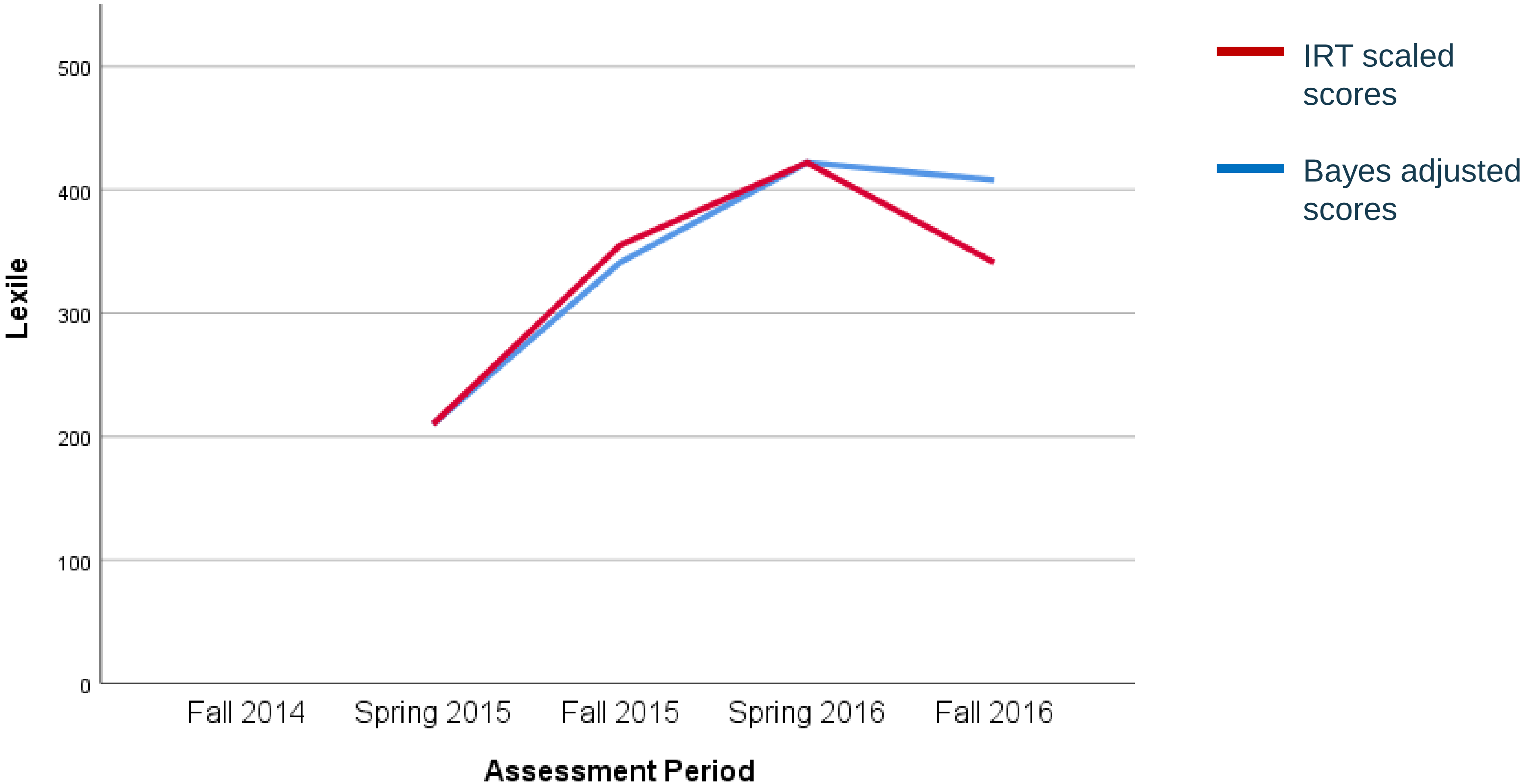
Example: Bayes “smooths out” a student’s measures



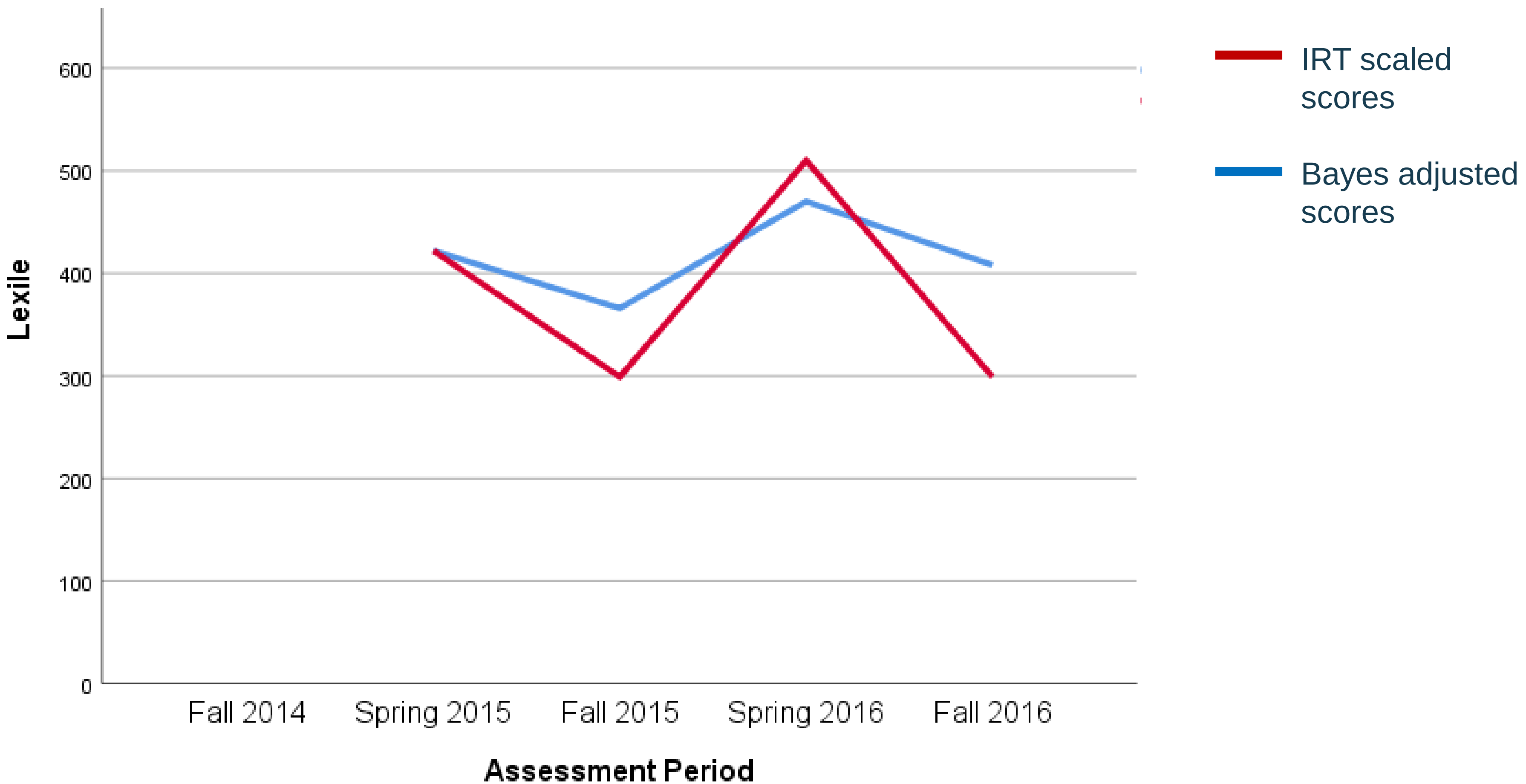
Example: Bayes makes very little difference



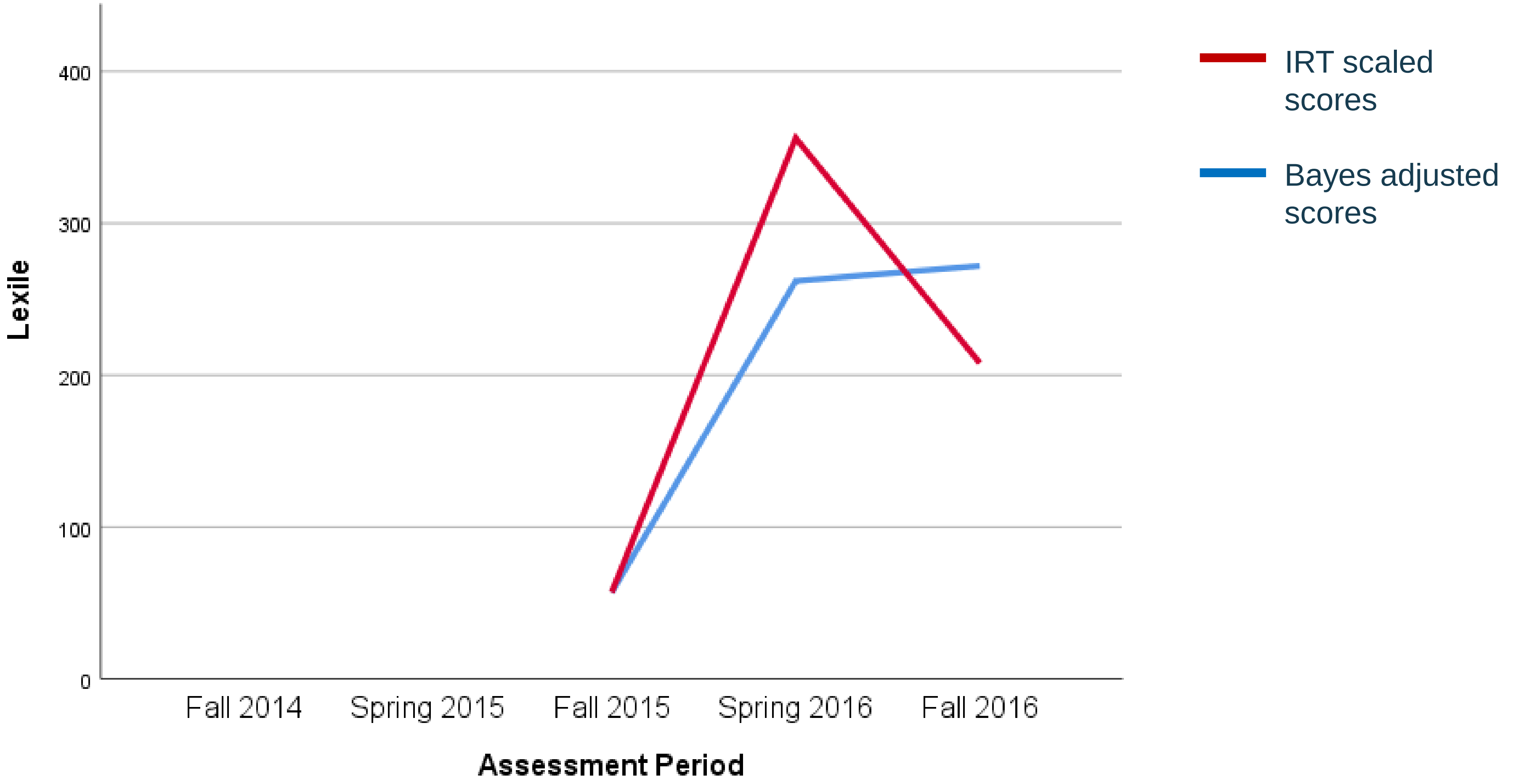
Example: Bayes reduces impact of a single poor test



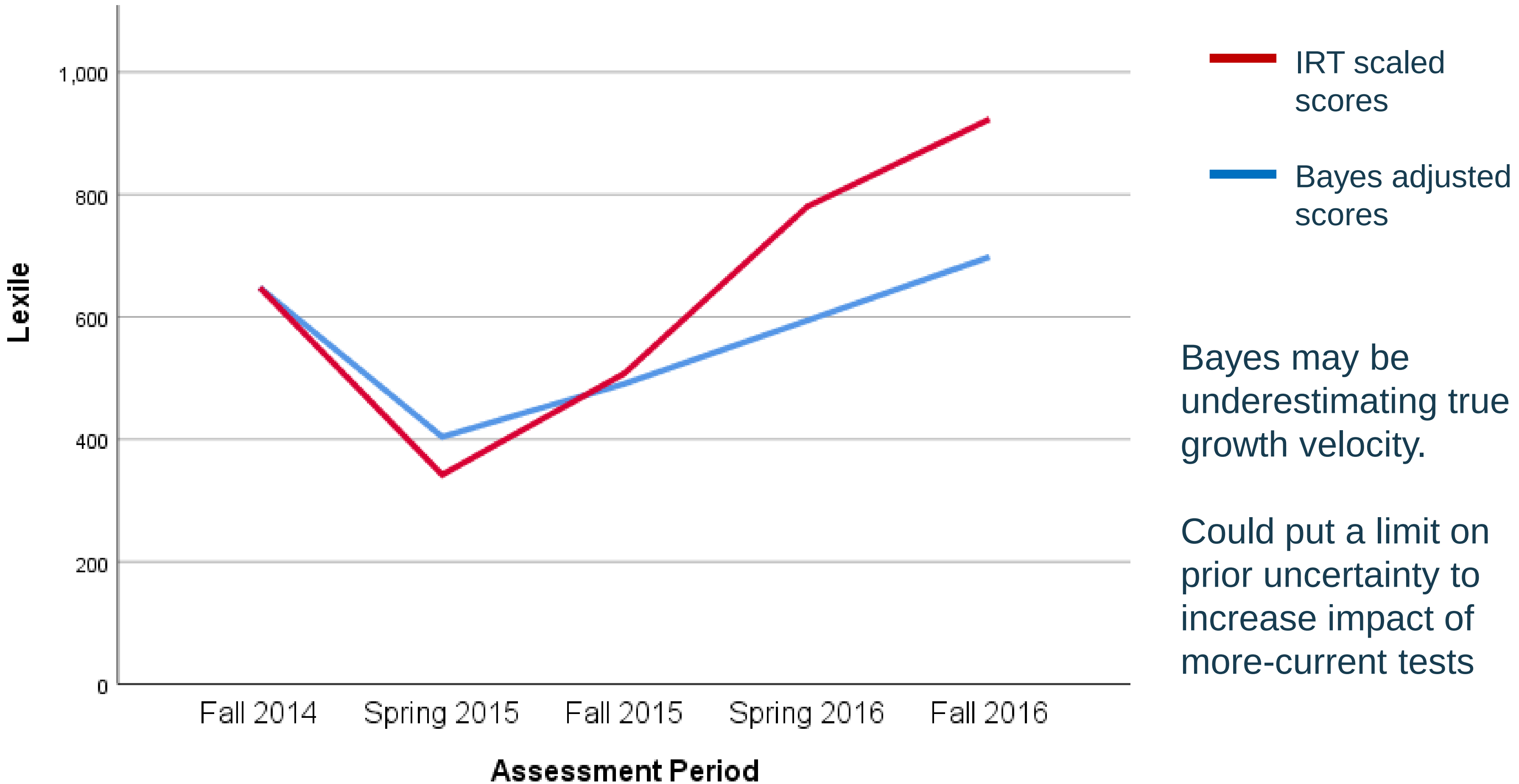
Example: Bayes reduces the peaks and valleys



Example: Bayes reduces the peaks and valleys



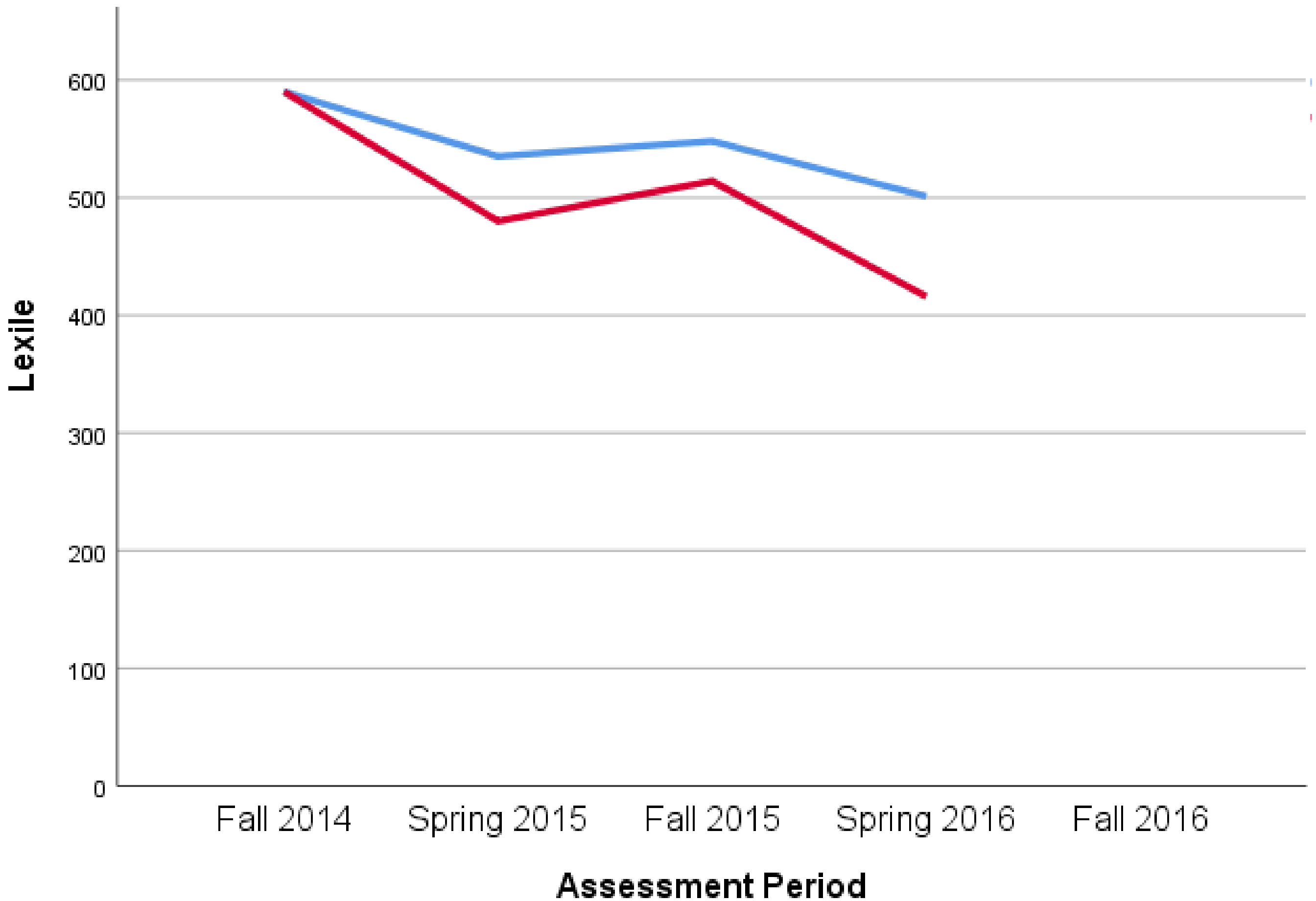
Example: Bayes reduces the peaks and valleys



Bayes may be underestimating true growth velocity.

Could put a limit on prior uncertainty to increase impact of more-current tests

Example: Bayes resists declining scores



IRT scaled scores

Bayes adjusted scores

Bayes may be overestimating true ability.

Research question

MetaMetrics.

To what extent do students' reading ability estimates differ when using Rasch-scaled scores from a single administration *versus* Bayesian scoring with priors?

Contexts:

- L2 students in an after-school English-language instructional program in S Korea (n=25)
- L1 elementary students in schools in the US following a curriculum-based reading program (n=20,928)

Data Source B: Group-Level Exploration

Assessment

- Progress monitor test associated with a reading program, administered 3 times during the school year
- Specific test forms developed for each grade

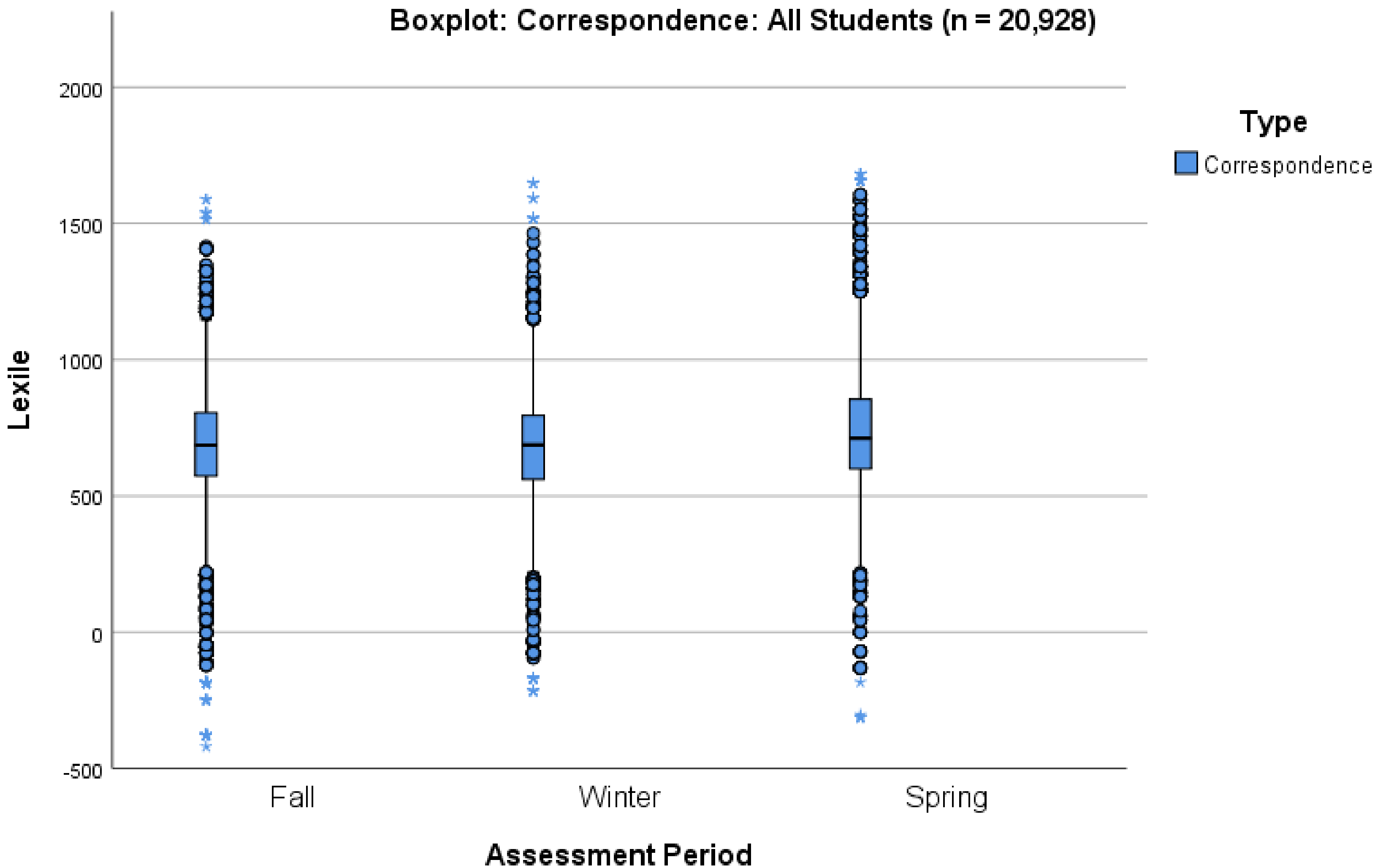
Grade	n
2	2
3	18
4	311
5	493
6	4,749
7	4,938
8	4,516
9	3,298
10	1,483
11	765
12	355
Overall	20,928

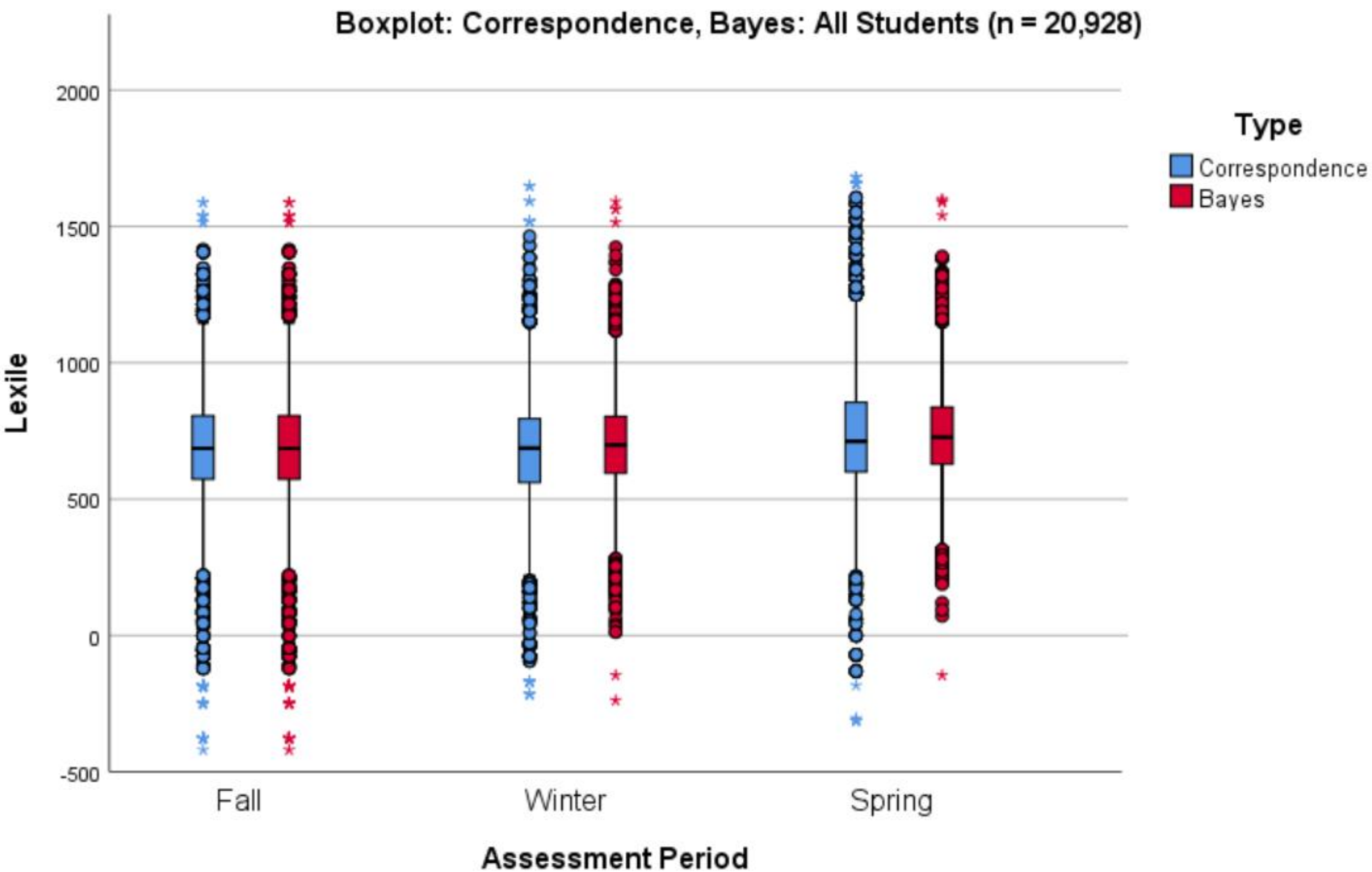
Sample

- US students in general education settings

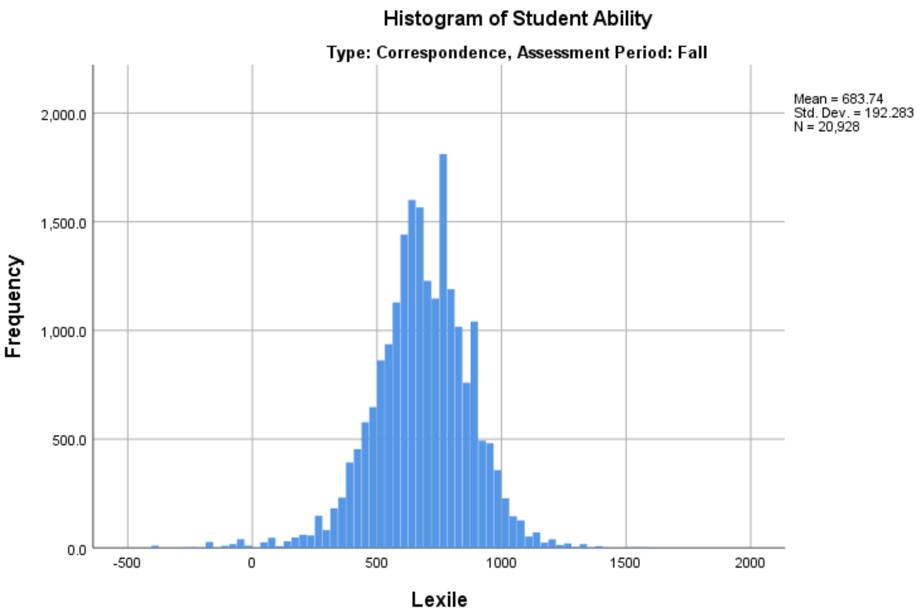
MetaMetrics.

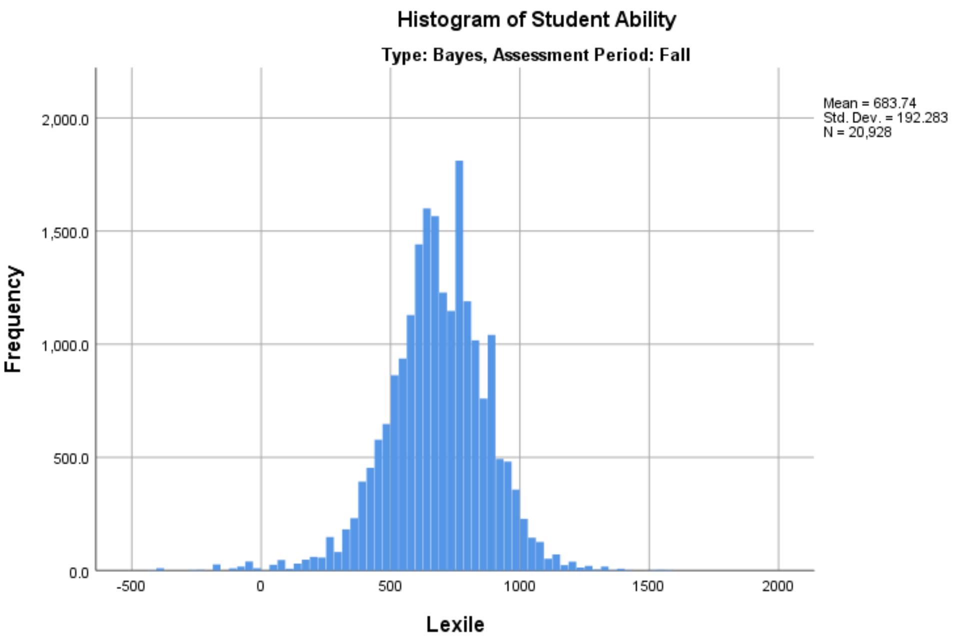
The following plots compare distributions of correspondence scores (i.e. non-Bayes) with Bayes adjusted scores.


Boxplot: Correspondence: All Students (n = 20,928)
Lexile
2000
1500
1000
500
0
-500
Fall
Winter
Spring
Assessment Period
Type
Correspondence


Boxplot: Correspondence, Bayes: All Students (n = 20,928)
Lexile
2000
1500
1000
500
0
-500
Fall
Winter
Spring
Assessment Period
Type
Correspondence
Bayes

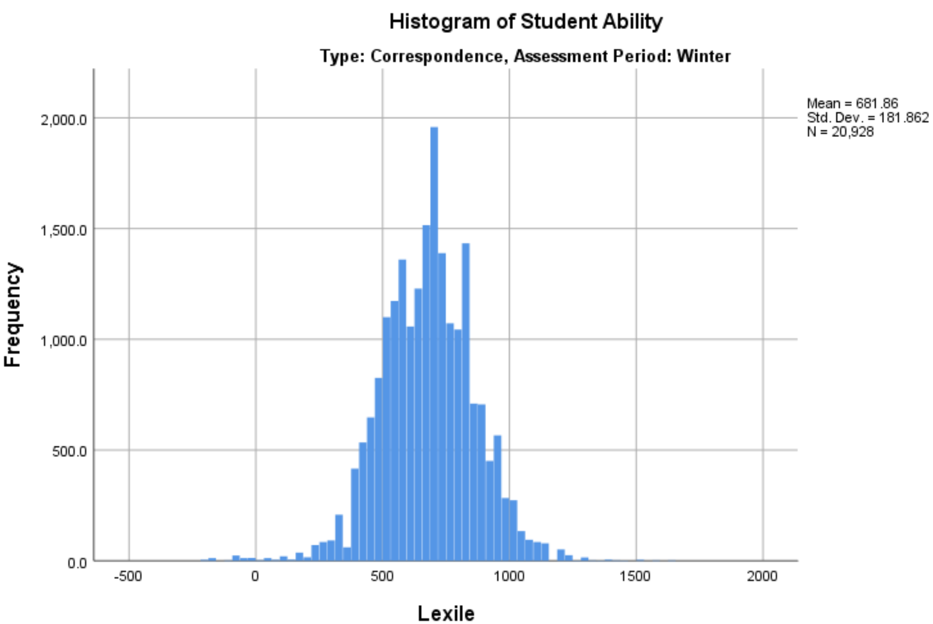
Correspondence vs Bayes, 1st Test Admin

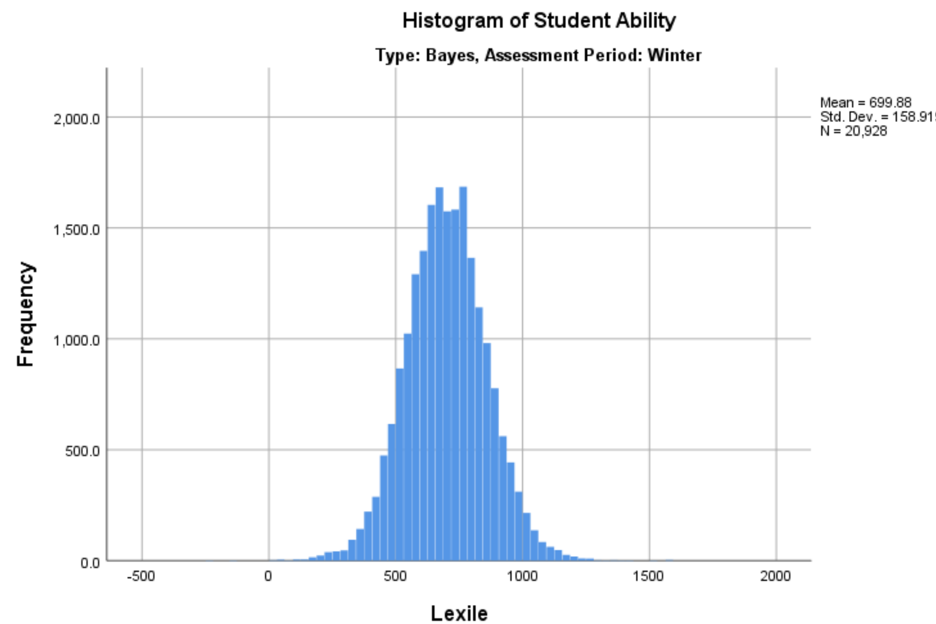




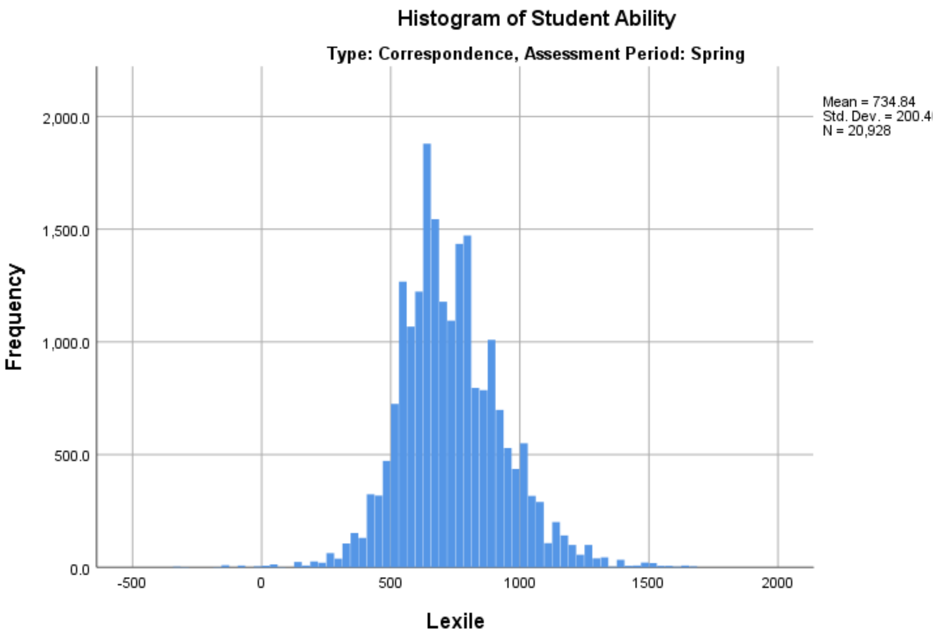
* Same starting point in the Fall

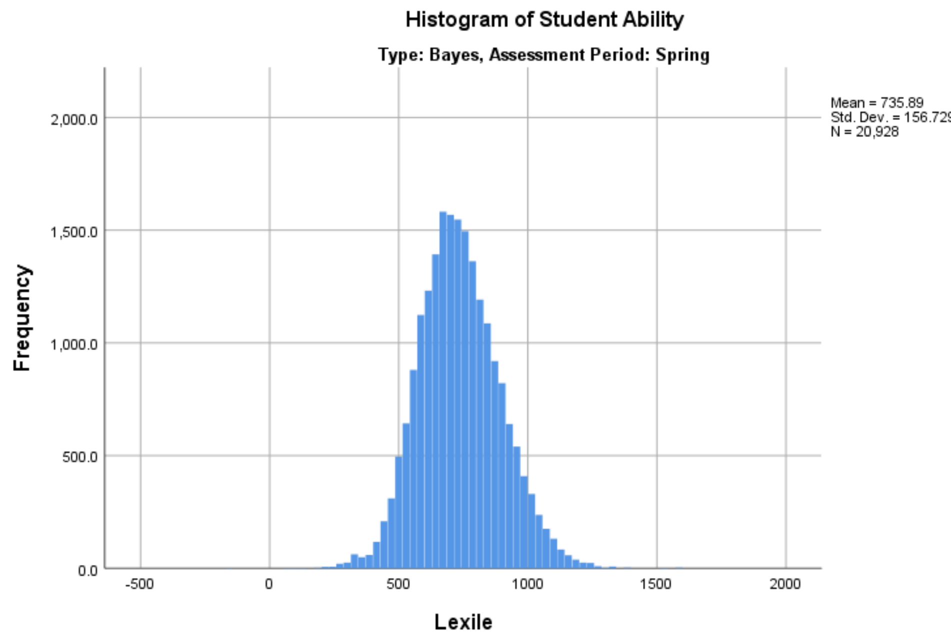
Correspondence vs Bayes, 2nd Test Admin





Correspondence vs Bayes, 3rd Test Admin





Bayesian Scoring in Repeated test-taking

MetaMetrics.

Conclusions from the data

- For individuals, Bayes scoring reduces the impact of a single poor or excellent test, and reduces the peaks and valleys (more in-line with expected student change)
- Over a large sample, it creates a more normalized distribution of test scores and reduces outliers
- Bayes scoring reduces growth or declines when they occur rapidly

Discussion

Advantages to this approach

- Uses all available data to inform ability estimates, not just the most recent observation
- Provides a more precise ability estimate
- Protects against extreme single performances (both poor and excellent)
- Is sensitive to time elapsed between tests
- Supports tunable features that can control how prior and current measures are used to produce an ability estimate

Discussion

Considerations

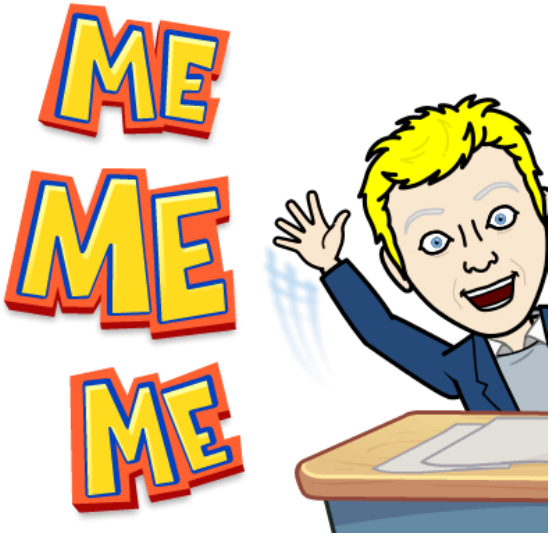
- Difficult to explain to people
- Suitable for when you want to understand “true skill”, not for performance instances (e.g. competitions)
- Tunable features require management and monitoring
  - For example: It is possible to underestimate growth if prior uncertainty is low, and prior proficiency estimates are weighted too highly

Discussion

Considerations

- People are used to receiving a test score; however, the notion of a latent ability estimate is more nuanced.
- We are no longer answering the question **“How did I do today?”;** rather, we are answering **“What is my updated ability estimate in light of today’s new information about me?”**
- Arguably, this is a better way to evaluate examinees, than to let them re-take a test numerous times until they are lucky enough to attain a high score

Questions



Alistair Van Moere
avanmoere@lexile.com
Lexile.co.uk