

Establishing baseline data for incidents of misconduct in the next generation assessment environment

Deborah J. Harris, Chi-Yu Huang, and Rya
Dunnington
ACT, Inc.

What (really) is benchmarking?

Benchmarking, a systematic comparison of the processes and practices of two or more companies or two or more units of a company, gauges the performance of an organization or unit relative to a peer.

(<http://hbswk.hbs.edu/archive/3746.html>)

Why is benchmarking “better” than statistical probability?

Consider looking at the responses two examinees have in common on a multiple choice test.

Statistically, one could argue that that the examinees have $\frac{1}{4} \times \frac{1}{4}$ chance of both selecting option “a” on item 1 randomly

BUT Are they selecting randomly?

Are they selecting independently?

Is option “a” more or less attractive?

Benchmark data provides a richer context in which to interpret results. It also serves as a way to “account for” variables and conditions we can’t really account for, or can’t account for easily.

For example:

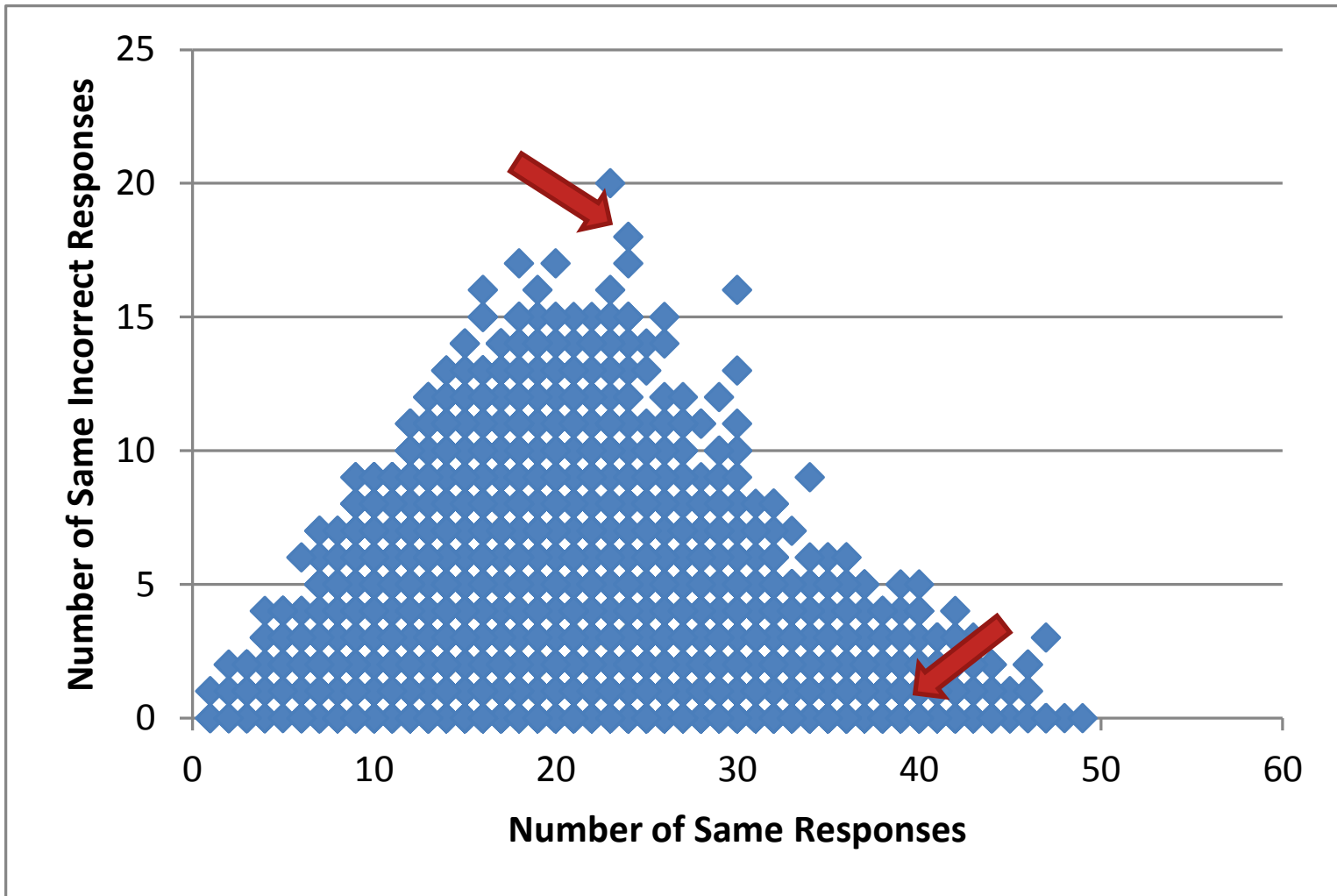
We “know” most examinees do not answer questions randomly. However, when we are looking at how similar two sets of examinee responses are, we don’t know what the tendency is for examinees who answer “c” to Item 7 to answer “d” to Item 8. Benchmark data can take this tendency into account, though indirectly.

Consider two pairs of examinees, one pair has 40 item responses in common; the other pair has 24 item responses in common.

We then look at additional information, such as how many of those common responses are incorrect.

Then we look at those pairs in relation to benchmark data.

Response Similarity



There are different ways to compute baseline data.

One might be considered, in a sense, a null case.

For example, for developing a copier set of baseline data, examinees who test in different states may be paired to determine how many like responses they have.

Other benchmark data might be considered, in a sense, less pure but more relevant.

For example, comparing all examinee pairs within a particular district, whether they tested in the same room or not, to have the “same curriculum” as a controlled factor.

We might initially run benchmark statistics by individual test form, or test date. But later, collapse across these and have a single set of benchmark data.

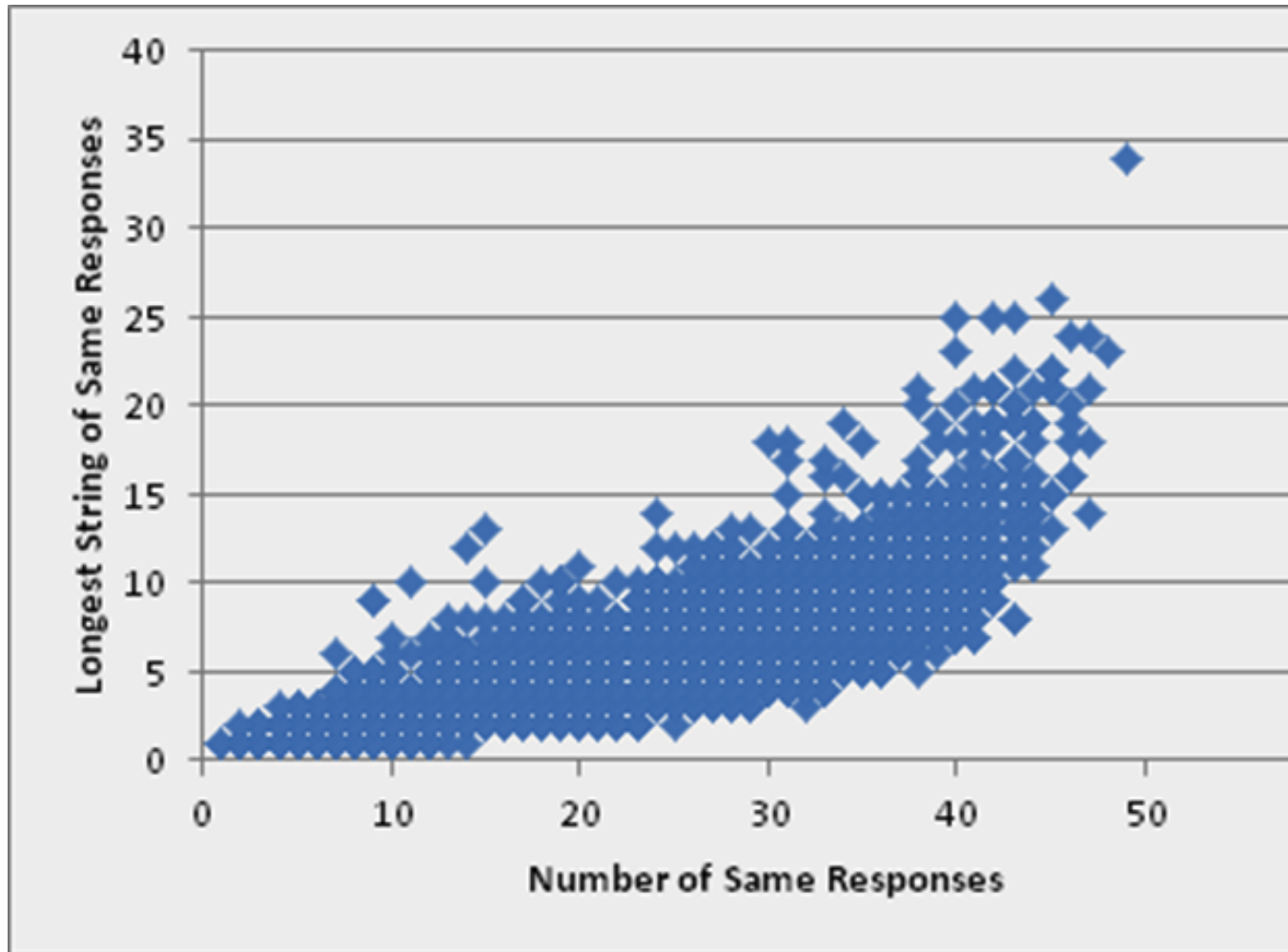
It's a balance between the context (this form, this set of co-examinees), and standardization (this is the single constant rule to flag any examinee , across forms and across examinee cohorts)

There is always someone most extreme.

There are always “outliers”.

The question is, does it matter in our context?

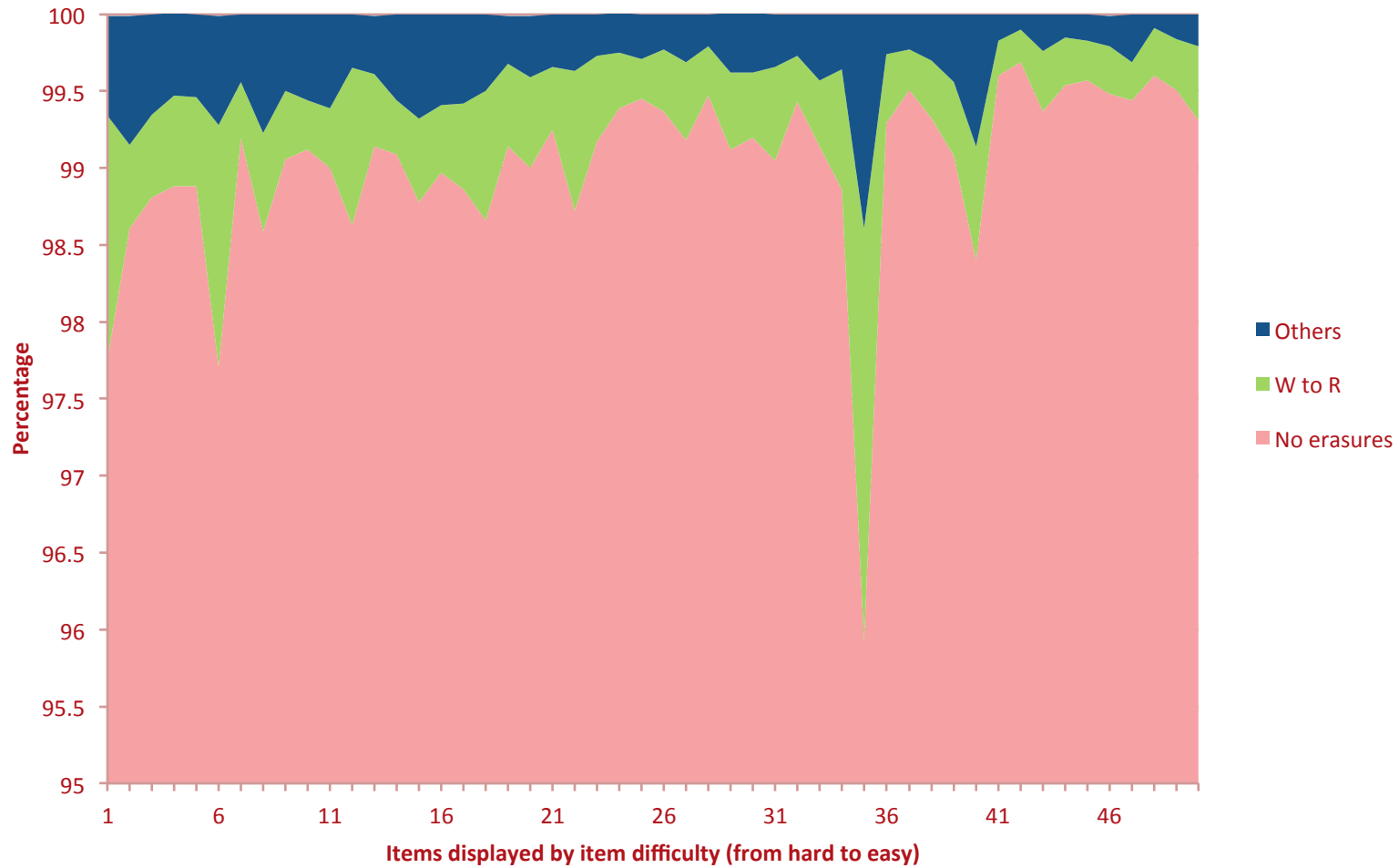
Response Similarity/String



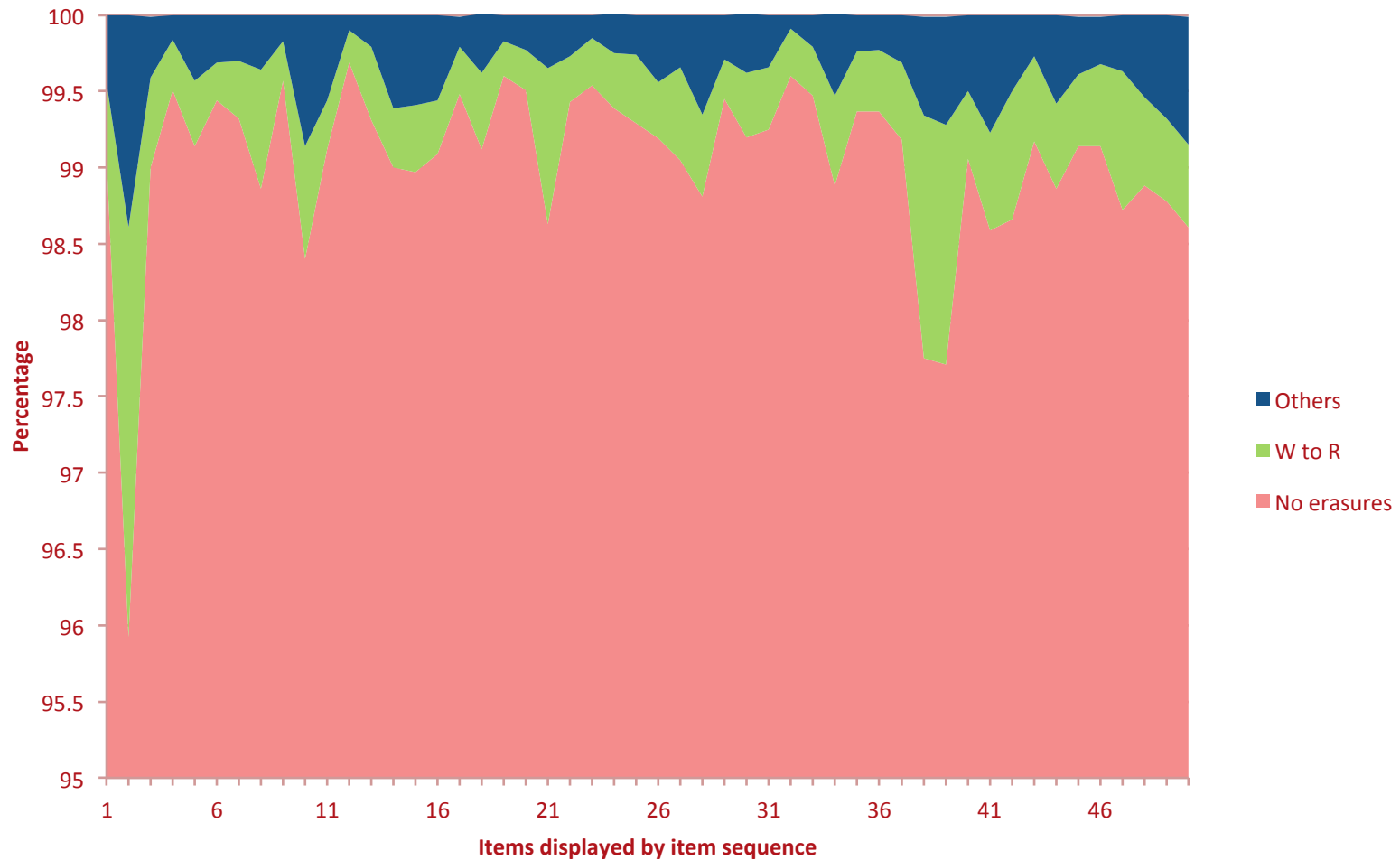
Having baseline data to compare values to both increases the likelihood of the values not being over or under interpreted, and provides a context in which to present information if the decision is made to pursue an incident.

But how we display the information may also matter.....

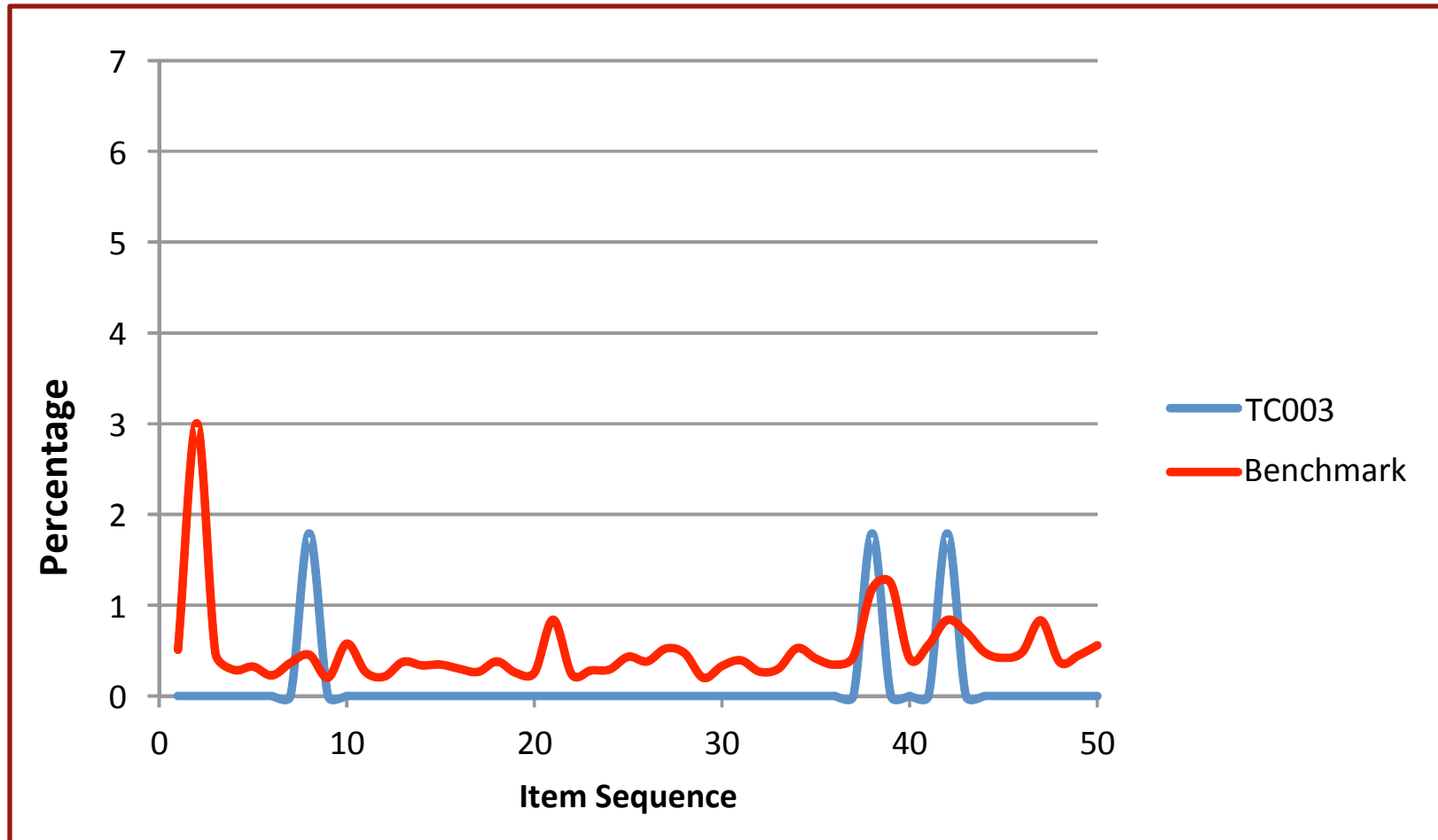
Erasure Patterns



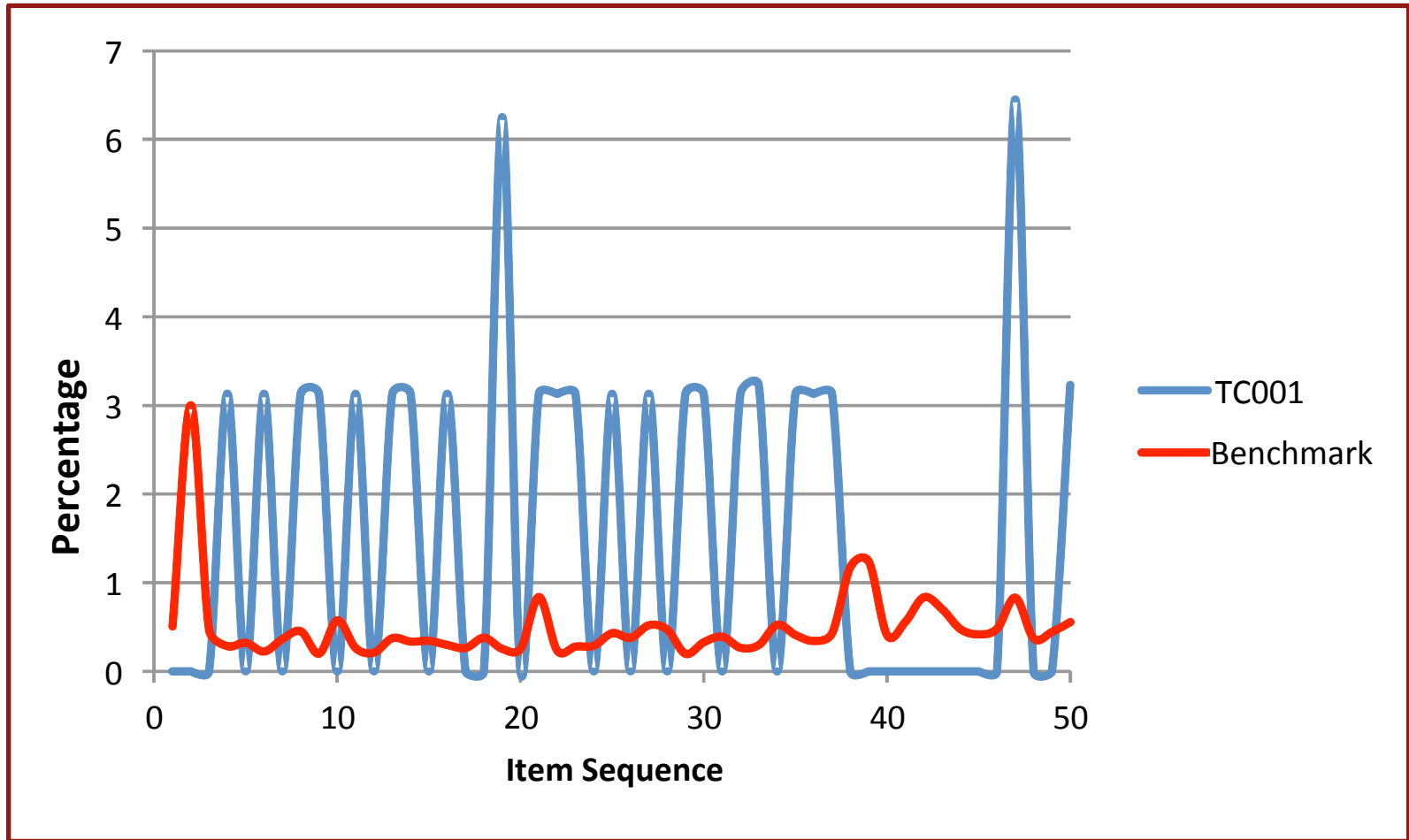
Erasure Patterns



Erasure Patterns: Wrong to Right

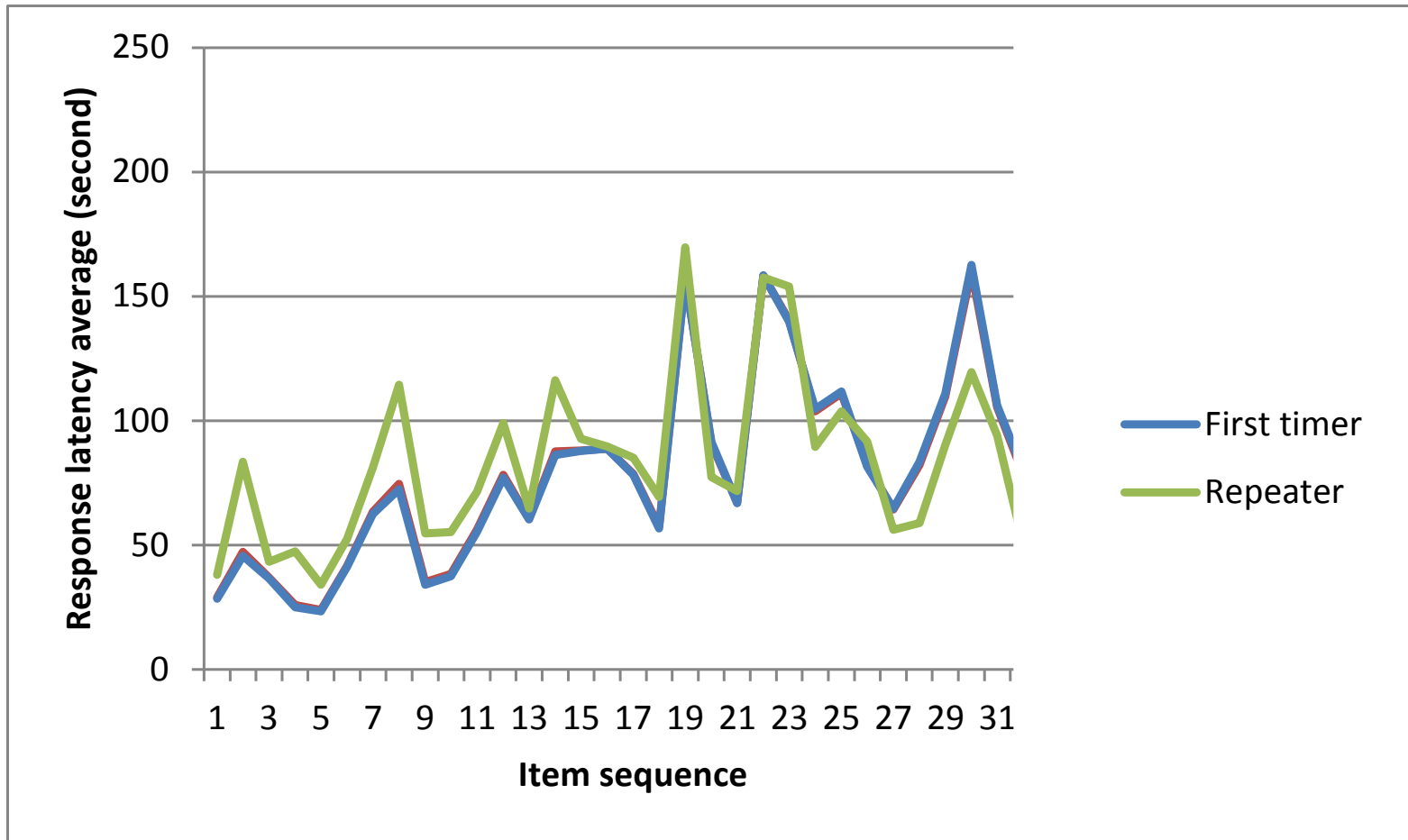


Erasure Patterns: Wrong to Right

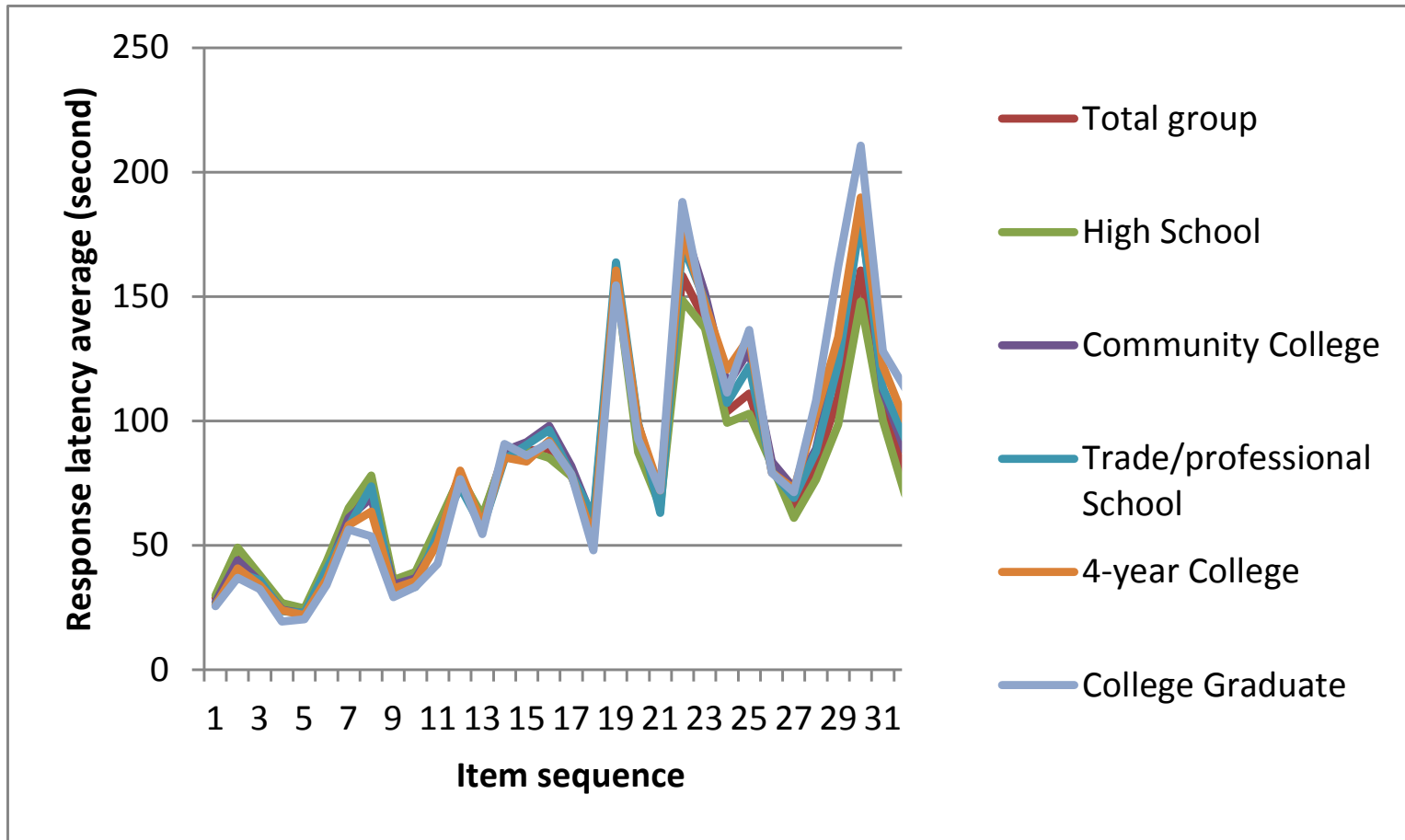


As stated earlier, benchmark data can be “refined”, or made more relevant....

Latency Data



Latency Data

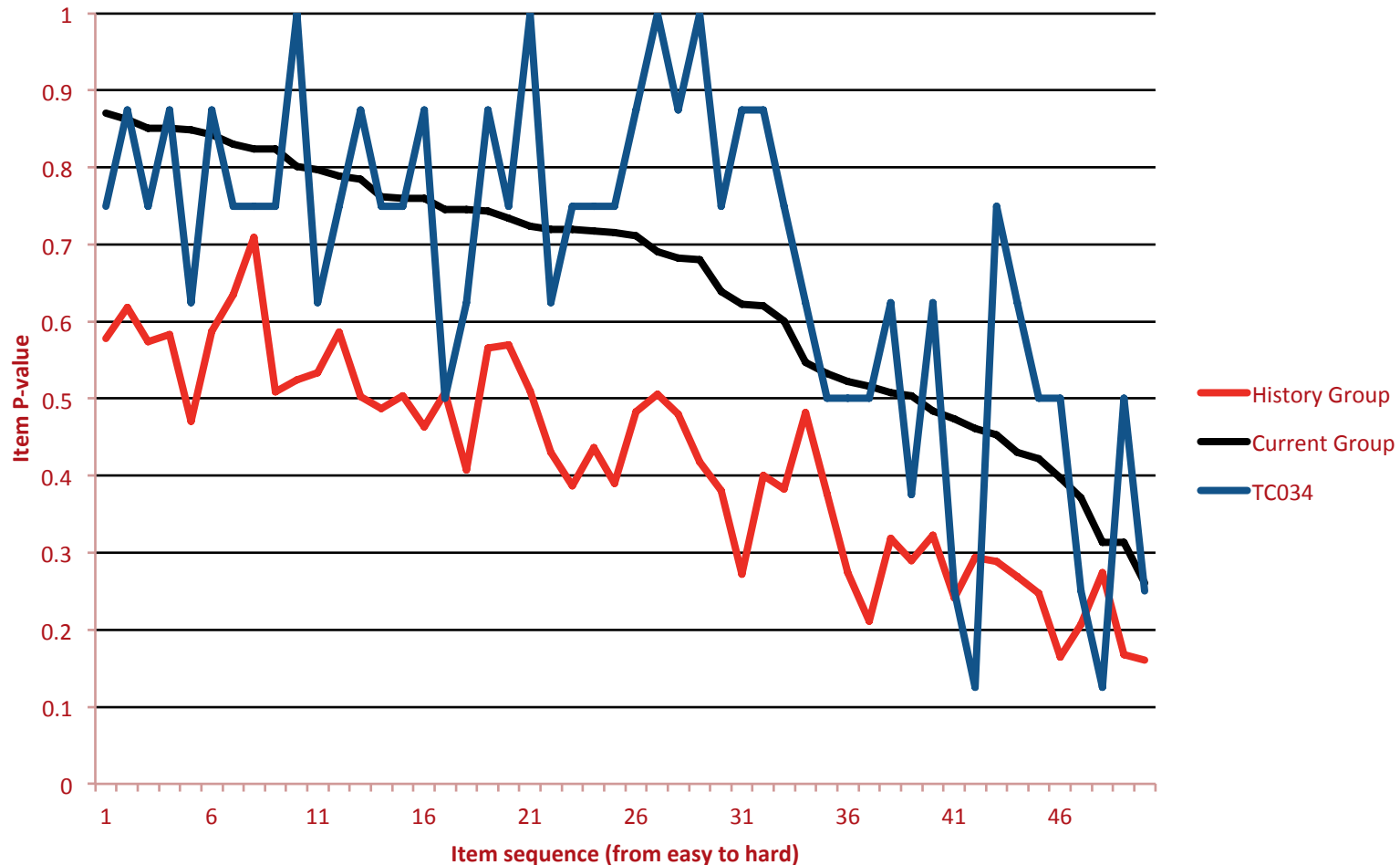


Option analysis (what percent of students choose each available option on a multiple choice or selection-type item) is often done to identify test site anomalies

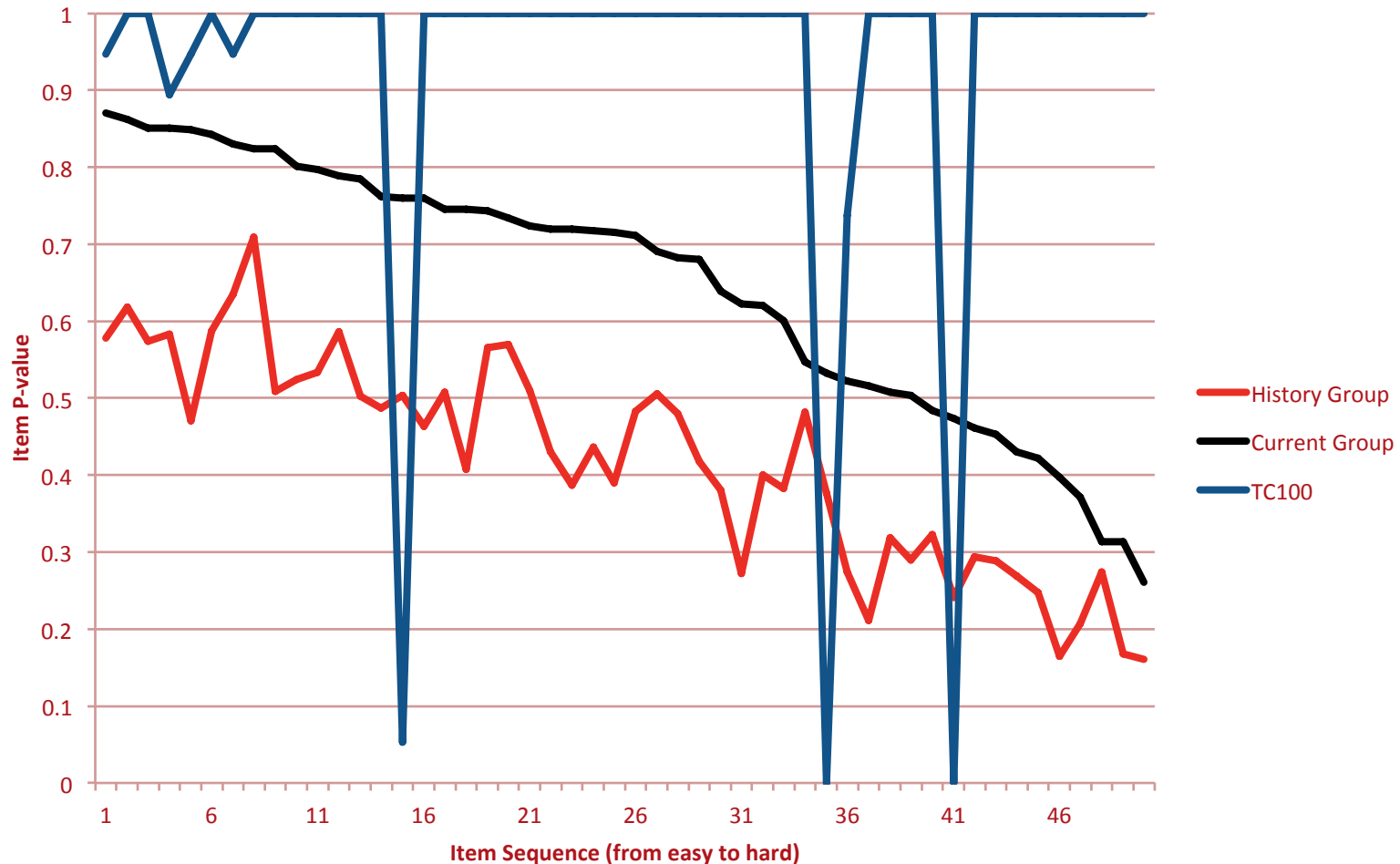
Item Performance by Test Center

TC	option1_item1	option2_item1	option3_item1	option4_item1	option1_item2	option2_item2	option3_item2	option4_item2
001	0	55.56	44.44	0	0	0	0	100
002	0	16.67	83.33	0	33.33	0	0	66.67
003	0	0	100	0	0	0	0	100
004	0	0	100	0	0	0	16.67	83.33
005	0	14.29	85.71	0	0	0	0	100
006	0	27.27	72.73	0	9.09	0	0	90.91
007	3.77	11.32	84.91	0	1.89	0	5.66	92.45
008	0	32	64	4	4	0	12	84

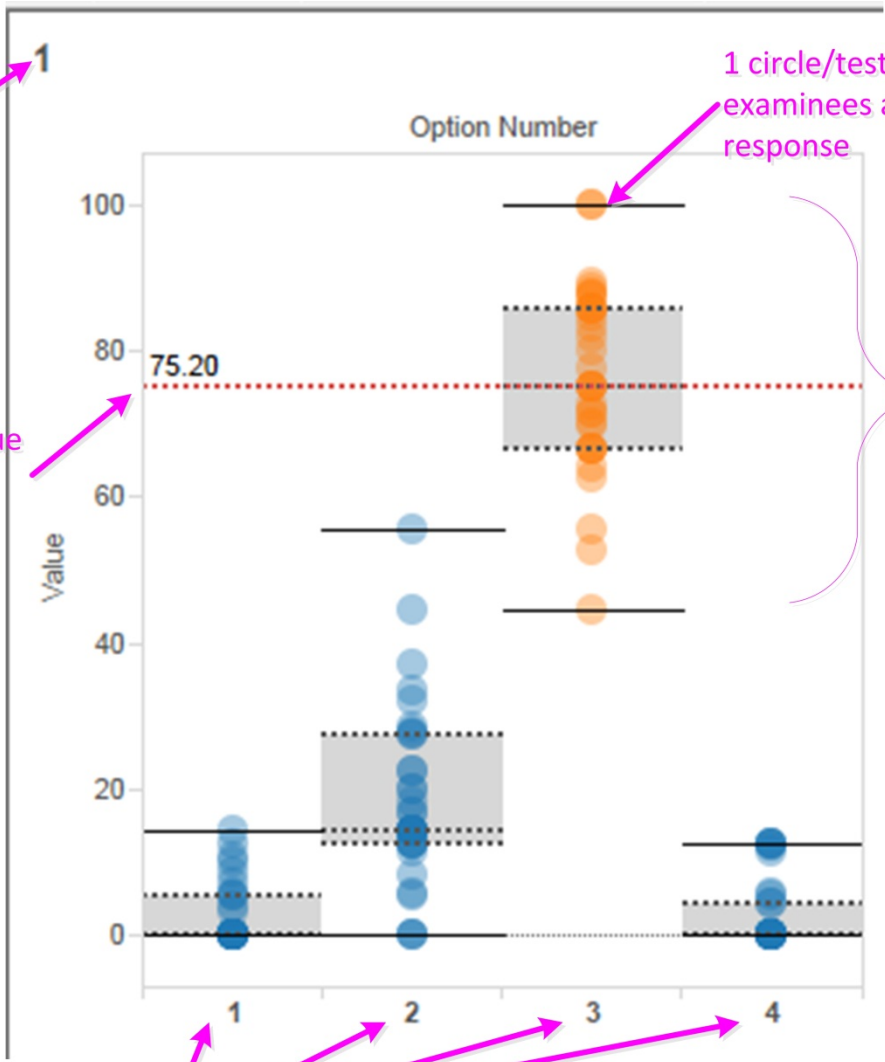
Option Analysis: Correct Option



Option Analysis: Correct Option

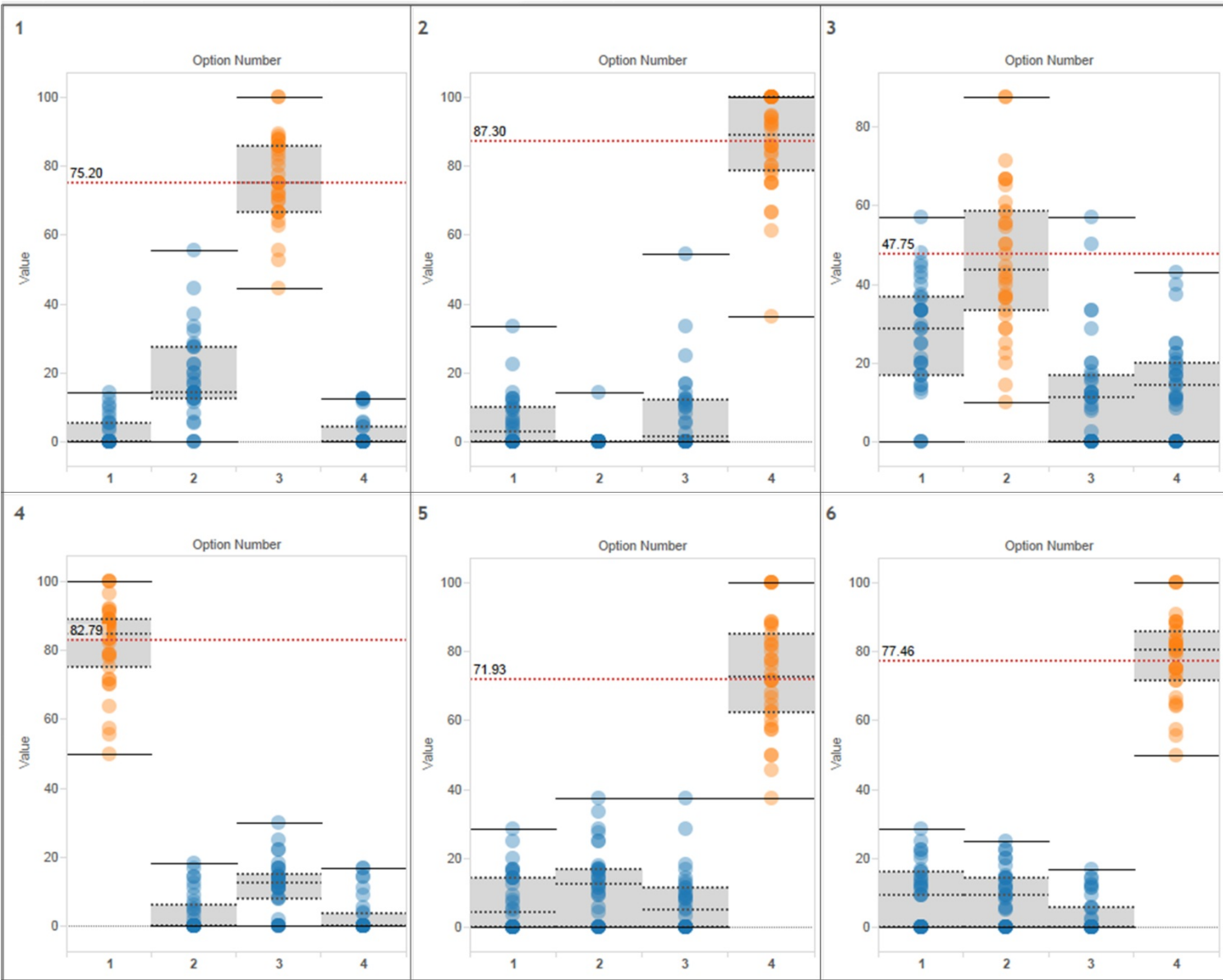


Item Number

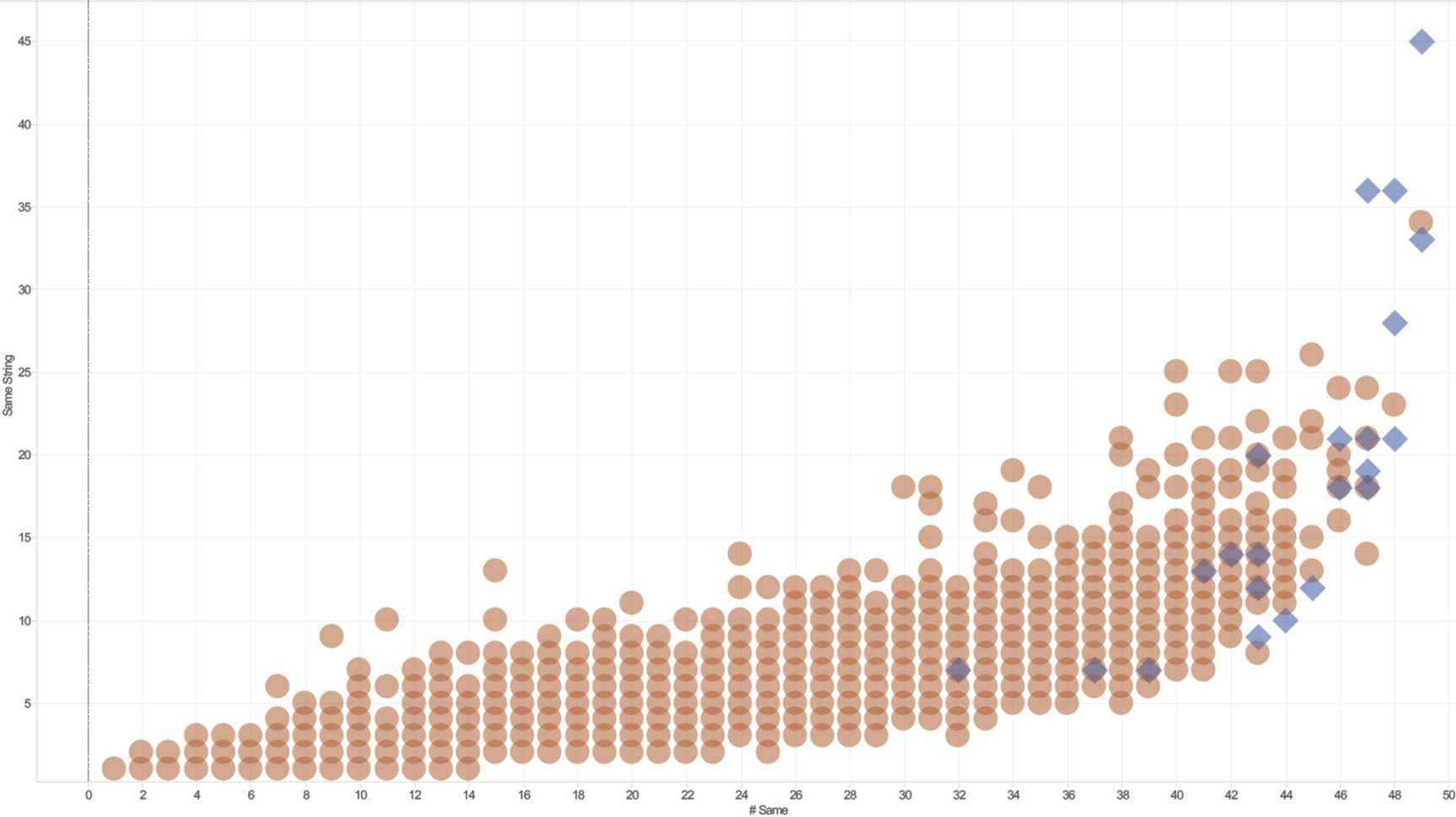


representation of P-value
(expressed as 0-100,
e.g.,
75.20 = p value .752)

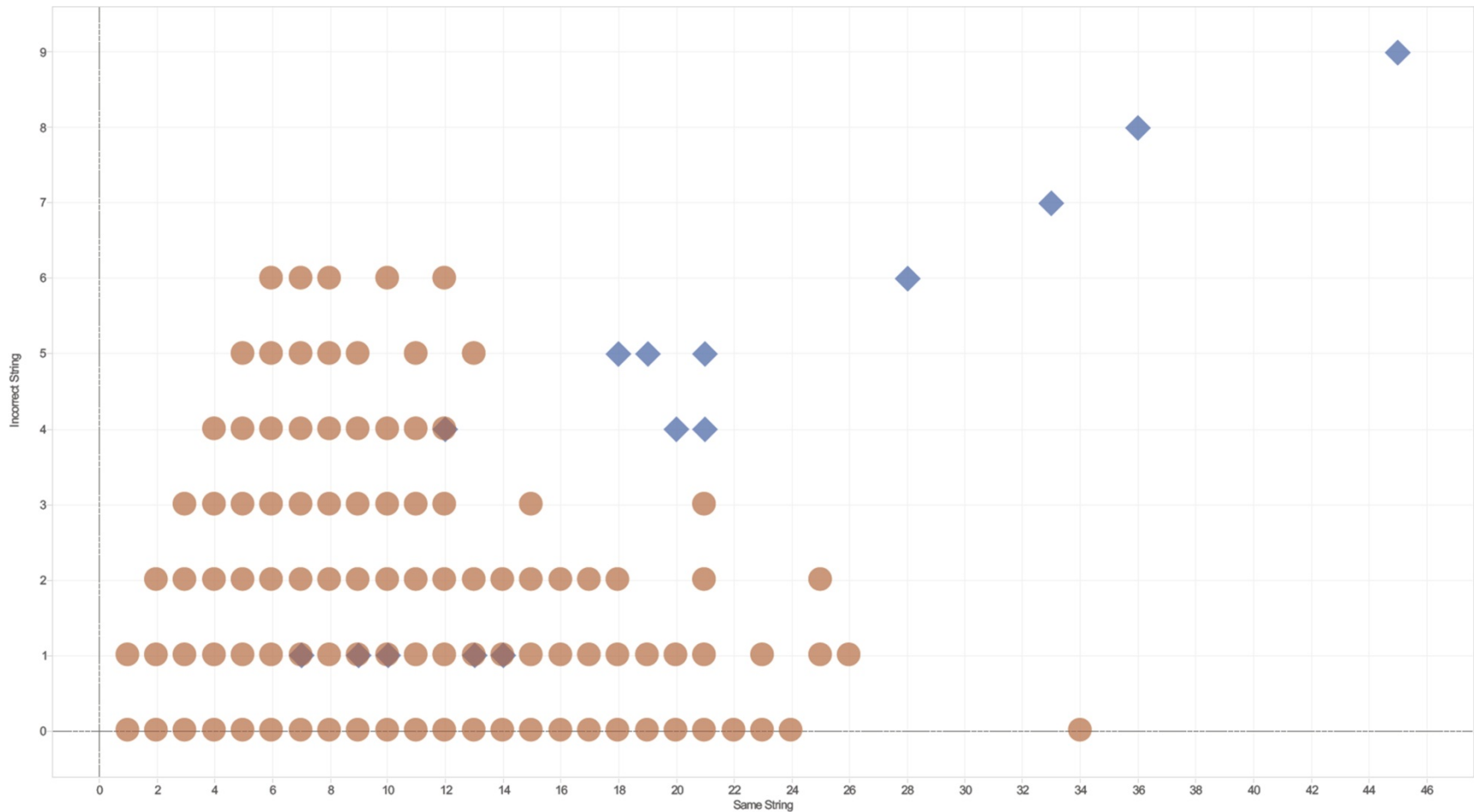
Responses



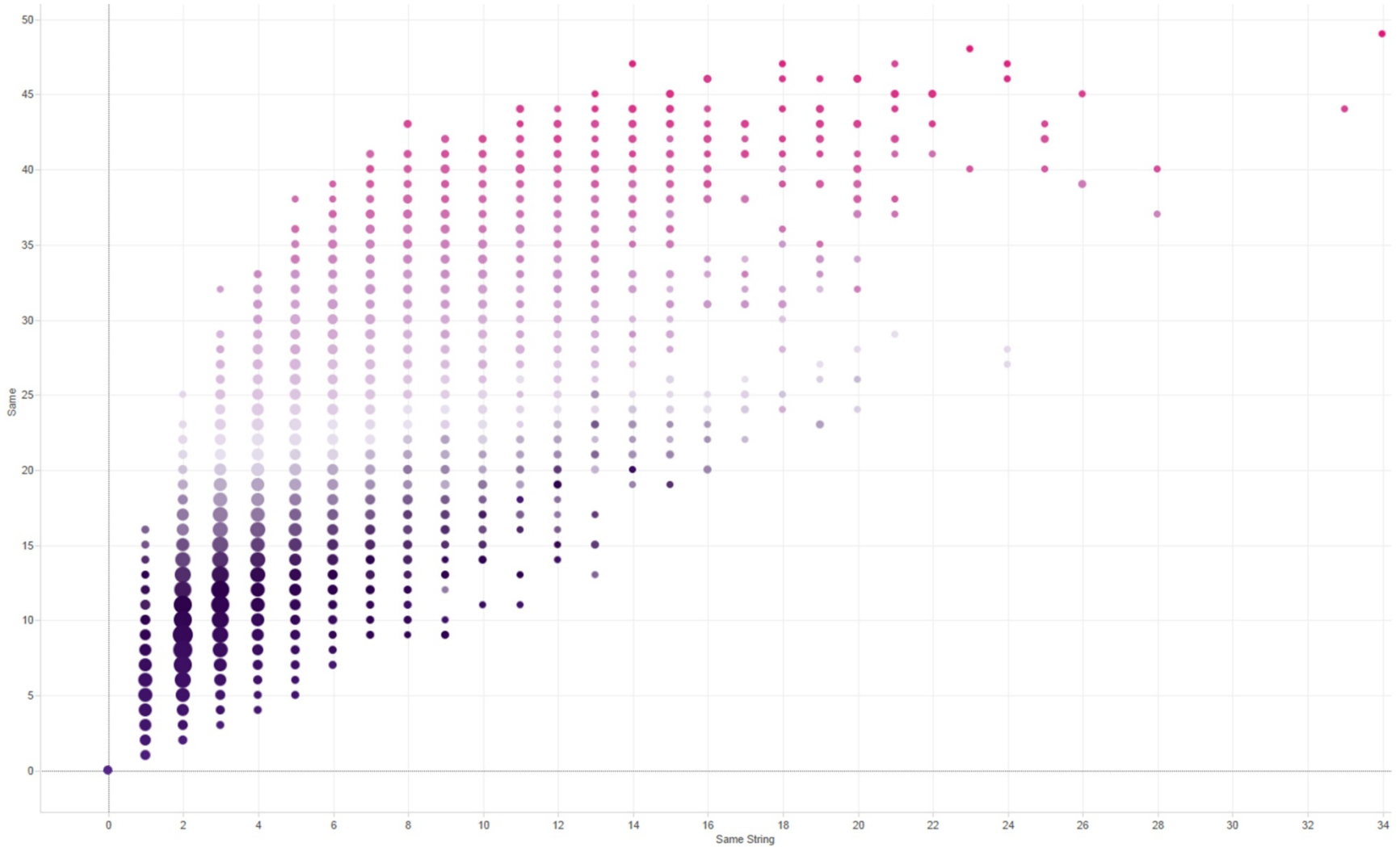
Number of Same Responses v. Longest Consecutive String of Same Responses




Longest Consecutive String of Same Responses v. Number of Incorrect Responses in String



Longest Consecutive String of Same Responses v. Number of Same Responses



Average Raw Score
3.50  49.50

Thank you