

# Designing an Assessment System

Richard P. Phelps

International Research-to-Practice Conference  
Nazarbayev Intellectual Schools AEO

Astana, Kazakhstan

October, 2016



“If a thing exists, it exists in some amount. If it exists in some amount, then it is capable of being measured.”

--René Descartes,  
*Principles of  
Philosophy*, 1664

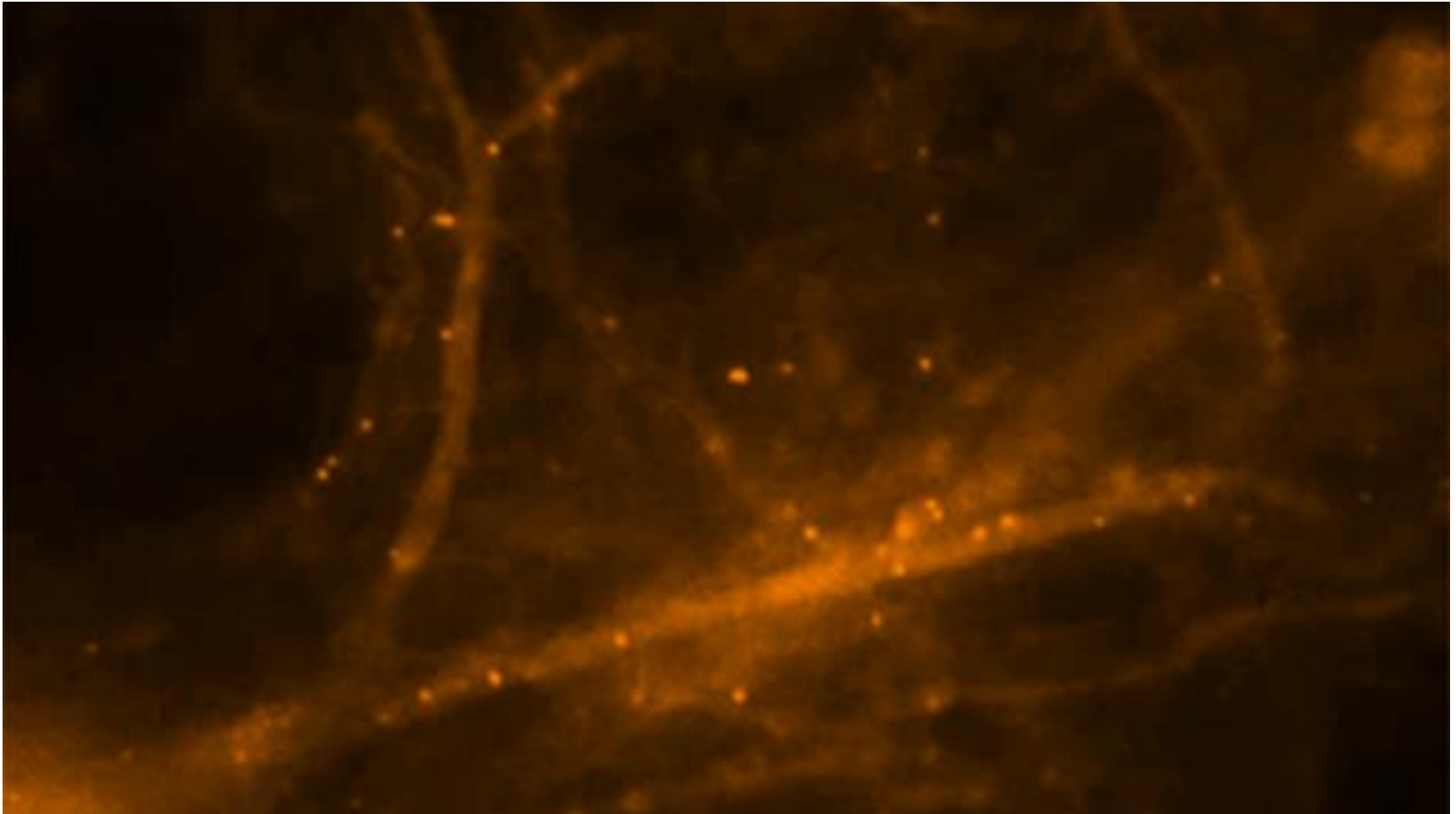
# Image of Protein Molecules Forming Memories

Albert Einstein College of Medicine, New York, January 2014

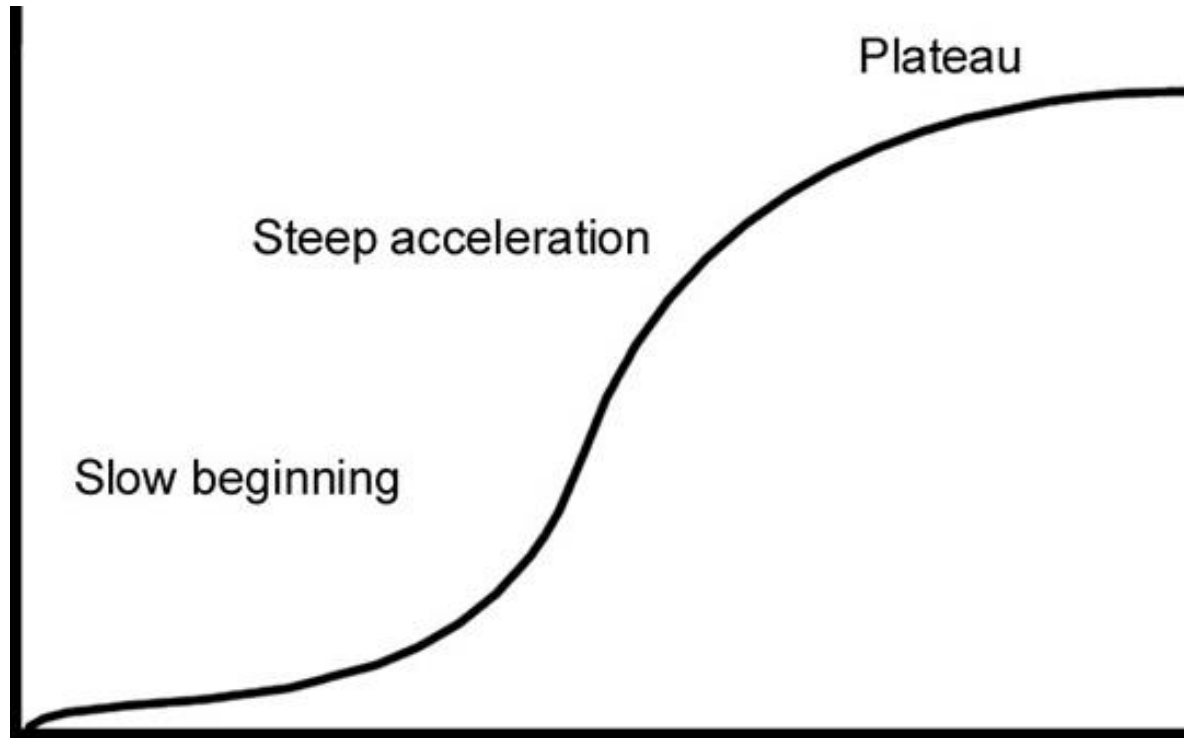


# Image of Protein Molecules Forming Memories

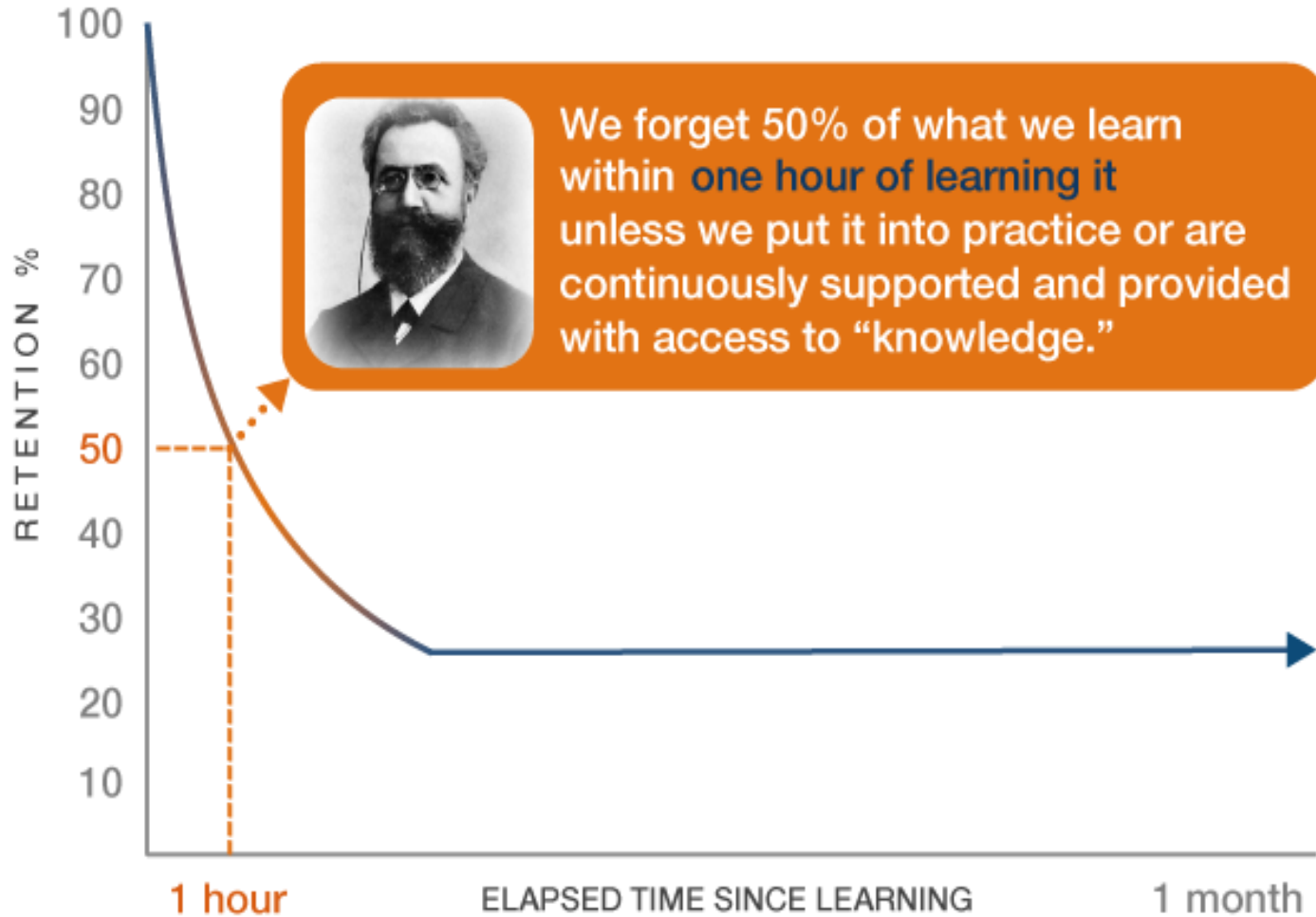
Albert Einstein College of Medicine, New York, January 2014



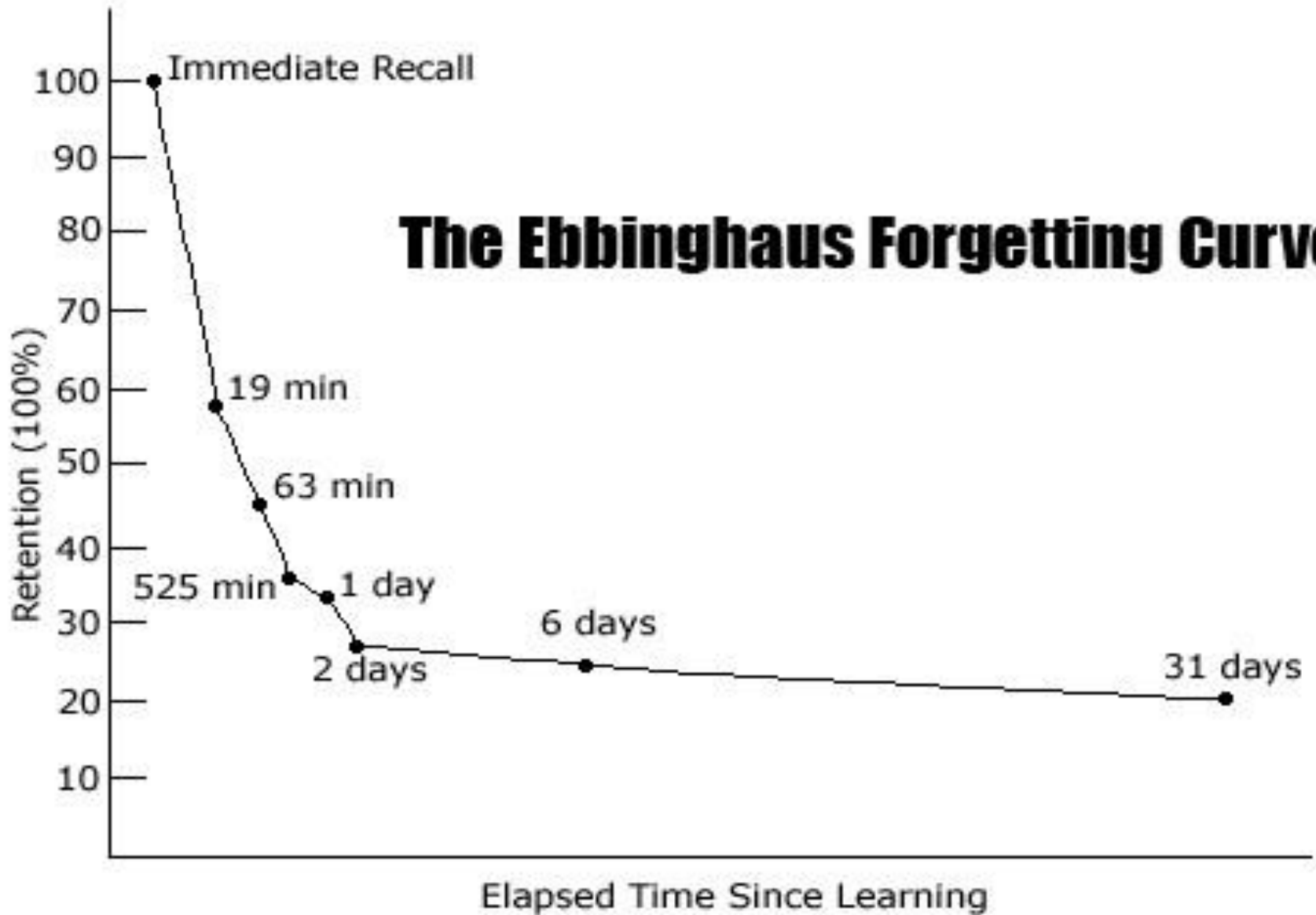
# Learning Curve



# Forgetting Curve (1870s)



# The Ebbinghaus Forgetting Curve



Ebbinghaus:

“Learning usually  
requires rehearsal  
or repetition”



Hermann Ebbinghaus  
(1850-1909)





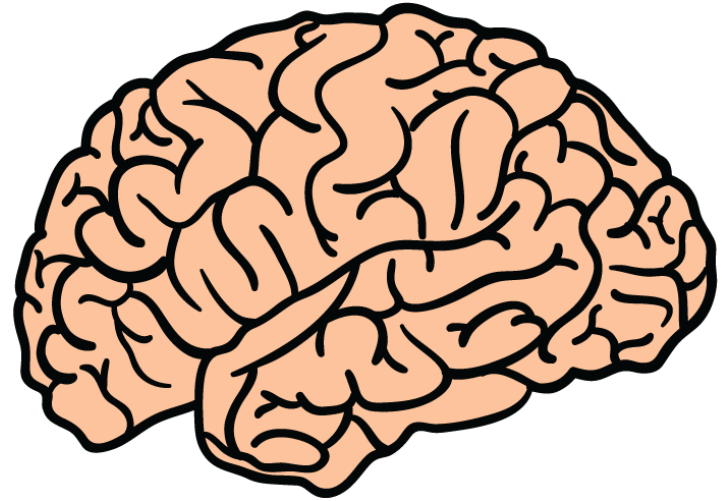
## Working Memory Capacity George Miller, 1950s

## Cognitive Load Theory John Sweller, 1980s



## Working Memory:

Ability to temporarily hold and manipulate information for cognitive tasks



Working Memory is challenged by:

new, unfamiliar information and

quantity of discrete bits of information

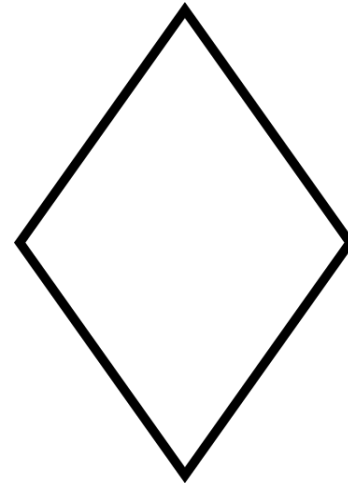
**I am thinking of a type of object, what is it?**

Description 1:

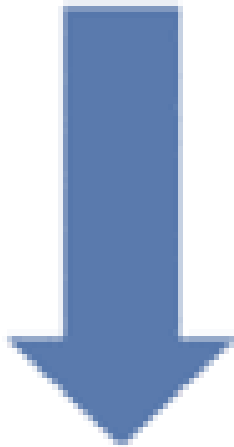
They are shapes, geometric plane figures, polygons, quadrilaterals, and parallelograms with opposite equal acute angles, opposite equal obtuse angles, and four equal sides

**I am thinking of a type of object, what is it?**

Description 2:



Intrinsic Load + Extraneous Load + Germane Load



**Manage**



**Minimize**



**Maximize**

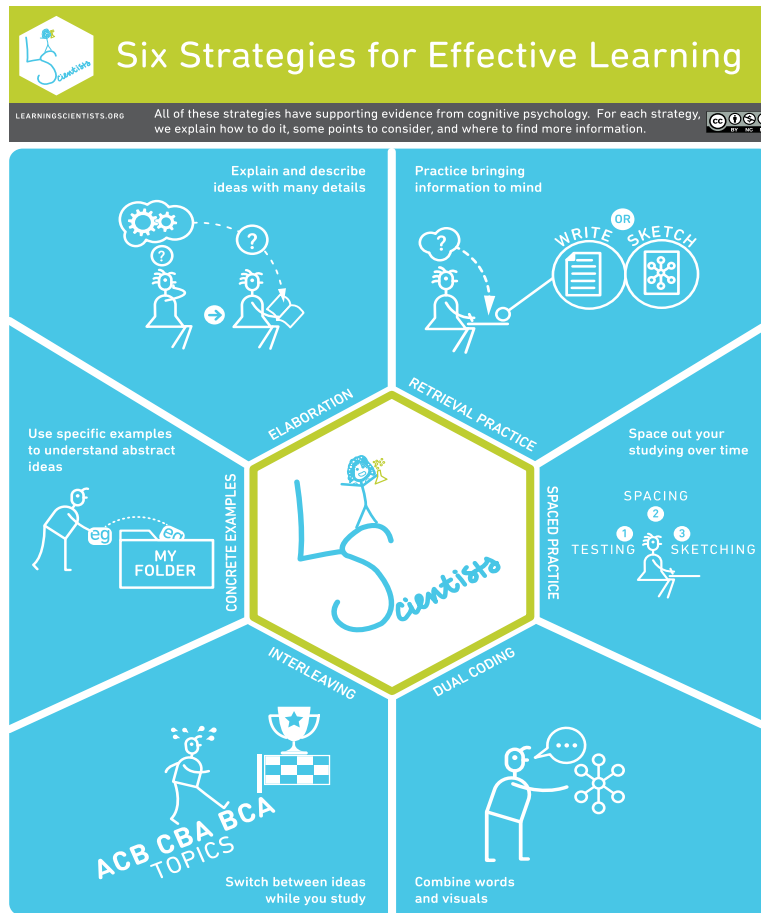
# Two centuries of research on learning concludes...



“...repeated retrieval during learning is the key to long-term retention.”

— Henry L. “Roddy” Roediger

# Cognitive Scientists' 6 Strategies for Effective Learning



Content by Yana Weinstein [University of Massachusetts Lowell] & Megan Smith [Rhode Island College] | Illustrations by Oliver Caviglioli [teachingow2s.com/cogsci]  
Funding provided by the APS Fund for Teaching and Public Understanding of Psychological Science

Retrieval Practice

Spaced Practice

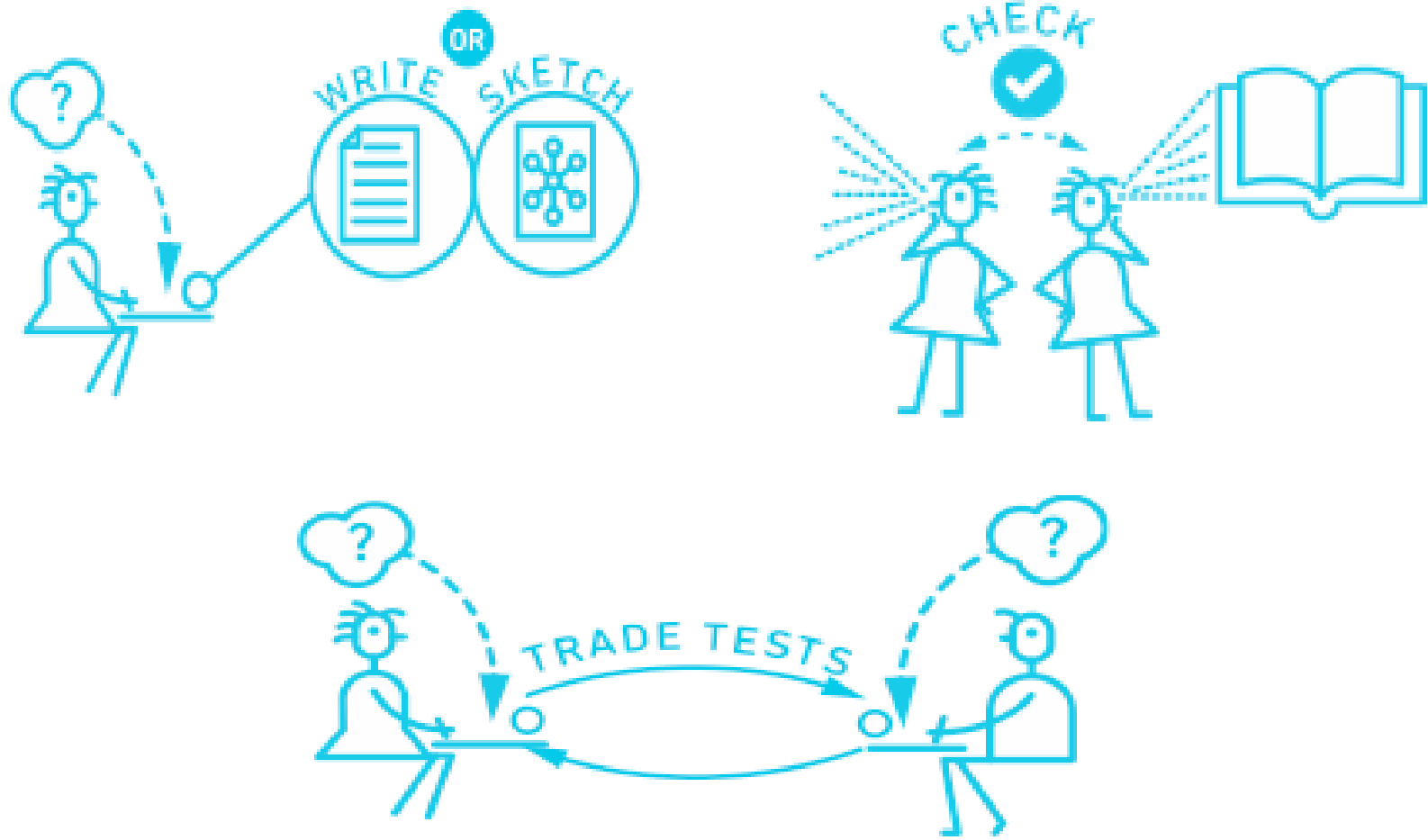
Dual Coding

Interleaving

Concrete Examples

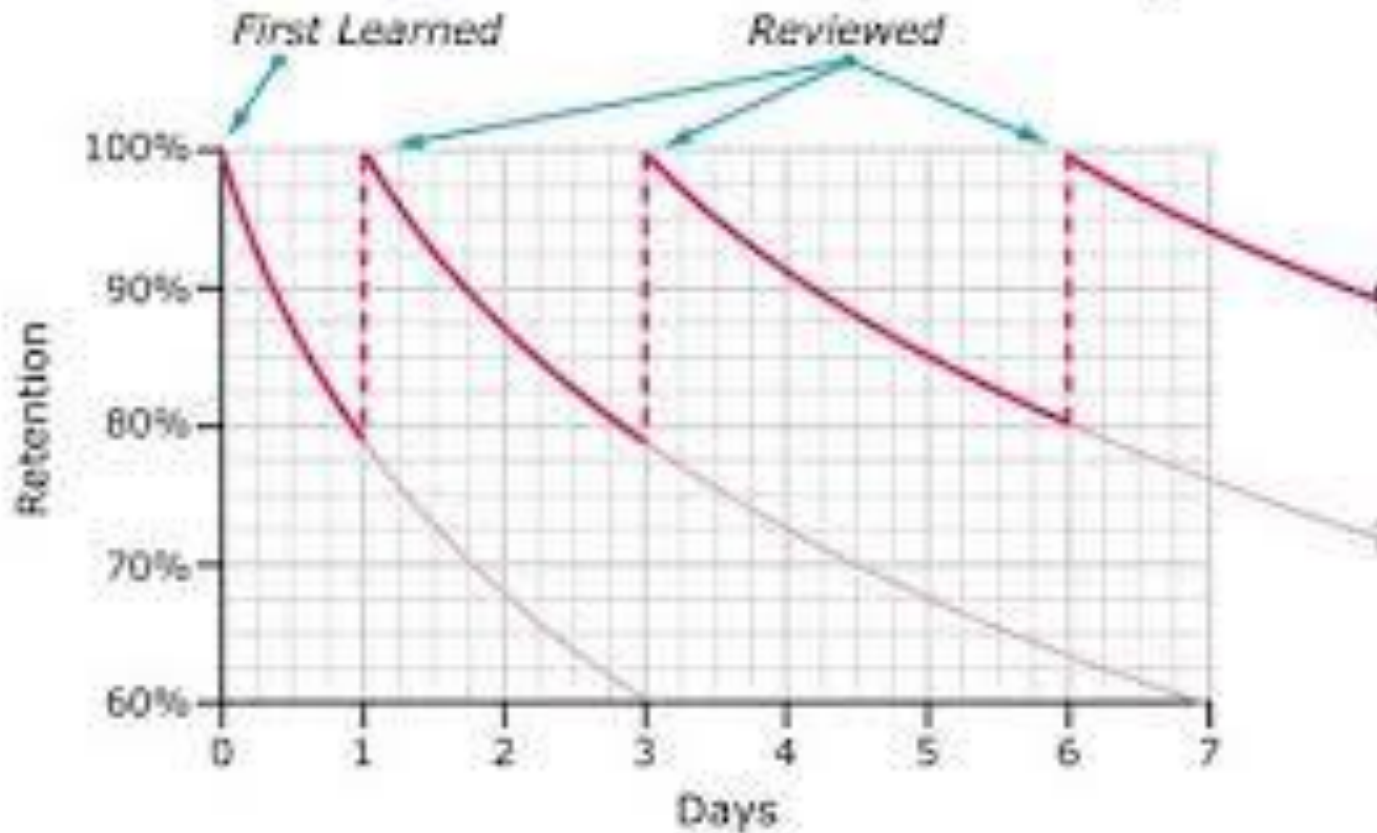
Elaboration

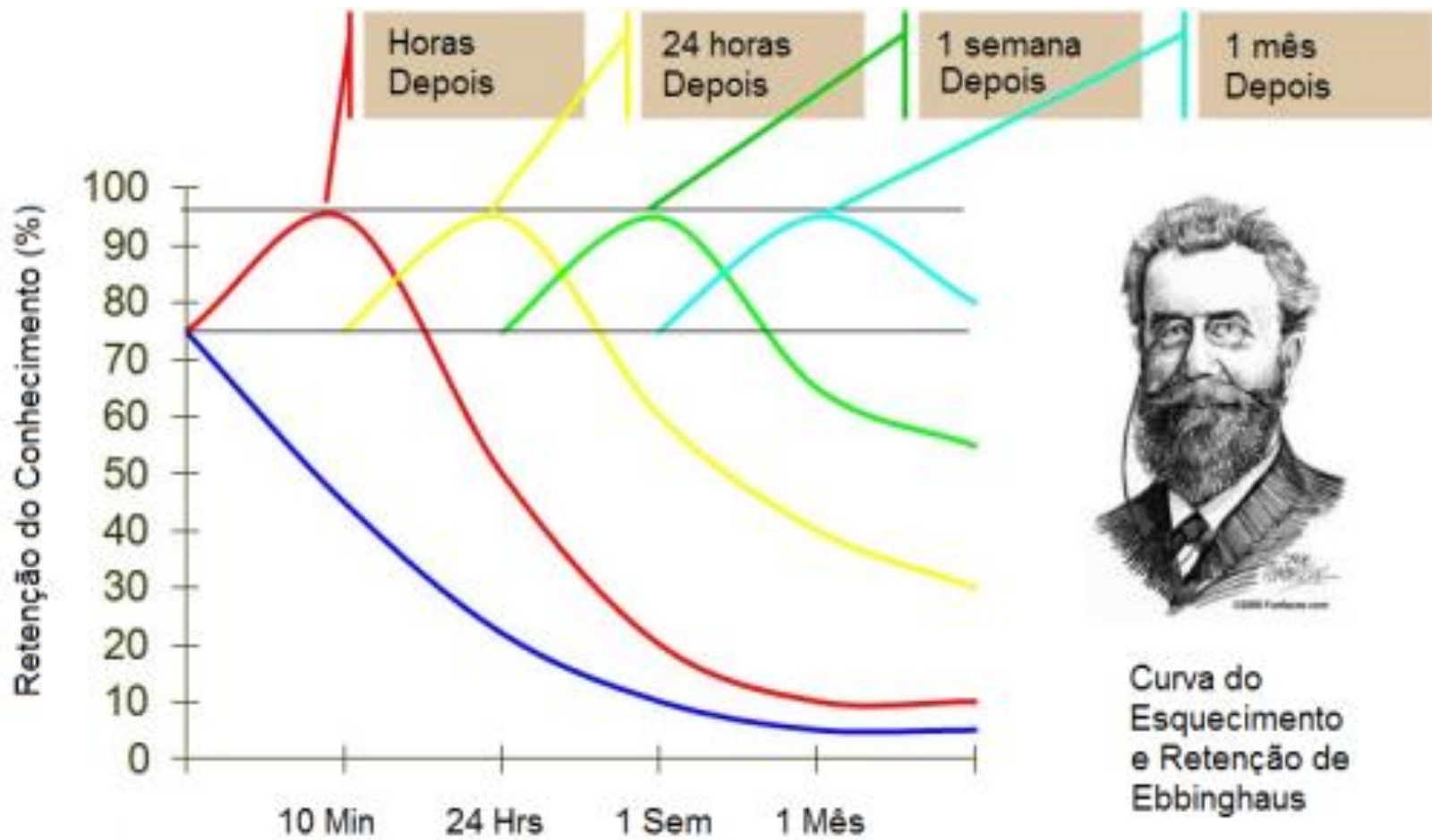
# Retrieval Practice





# The Forgetting Curve with Spaced Repetition





Curva do Esquecimento e Retenção de Ebbinghaus

# Implications for Teachers 1

Most teachers should test more frequently, ...with smaller, shorter, low-stakes tests



Understand that useful assessment can be short and simple.

## Implications for Teachers 2



Does the test format matter?

- multiple-choice?
- essay?
- short answer?
- oral?
- demonstration?
- ...etc.?

Not so much.

## Implications for Teachers 3

Tests provide feedback to teachers about what works and what does not



Just like students can learn by testing each other; teachers can help each other by reviewing each others' tests.

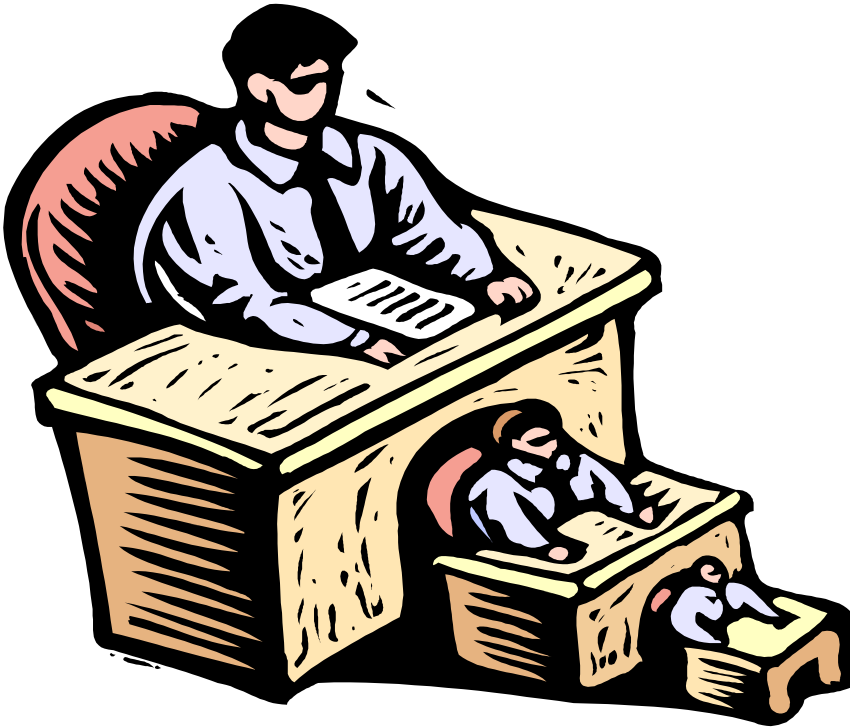
Cognitive Psychology  
experiments were  
conducted with  
“formative” tests in  
schools and classrooms

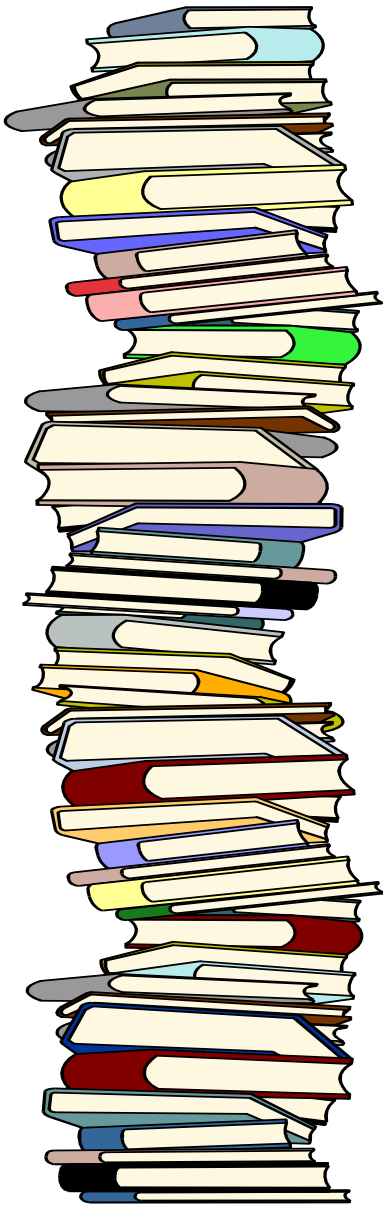


# What about systemwide, large-scale tests?

First priority:

do no harm to the  
formative testing  
programs in schools  
and classrooms





## The effect of testing on student learning

- 12-year study, read >3,000 documents
- analyzed close to 700 separate studies, and more than 1,600 separate effects
- 2,000 other studies were reviewed and found incomplete or inappropriate
- hundreds of other studies remain to be reviewed



# The effect of testing on student learning

245 Qualitative studies

813 Surveys or Polls

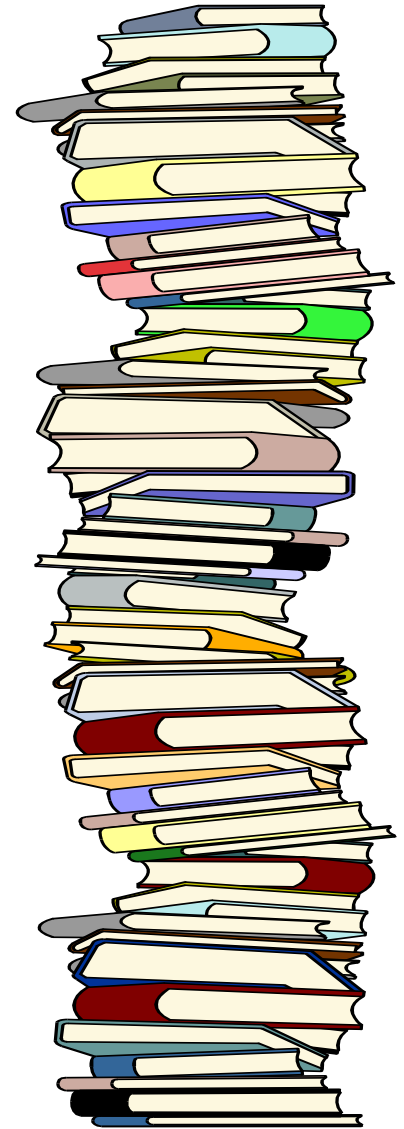
640 Quantitative Studies:

Experiments:

School- and classroom-level

Multivariate studies:

Large-scale testing programs



# Meta-analysis

A method for summarizing a large research literature, with a single, comparable measure.



( 0.5 effect size  $\approx$  1 grade level of learning )

# Findings from Phelps (2012):

- Survey study effect sizes average  $>1.0$
- Over 90% of qualitative studies positive
- For quantitative studies, univariate effect sizes positive and stronger when:
  - Testing more frequently
  - Testing with feedback
  - Testing with stakes



# Findings from Phelps & Silva (2015)

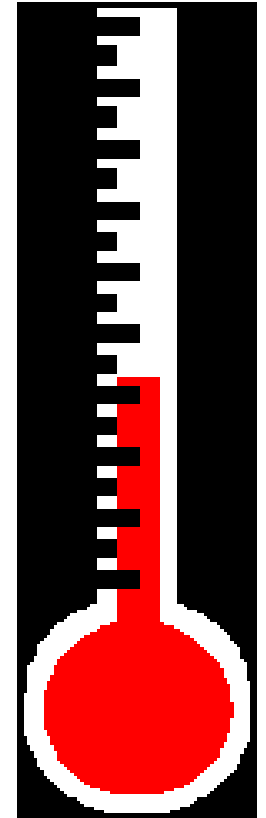
For quantitative studies, effect sizes vary  
between 0.55 and 0.88:



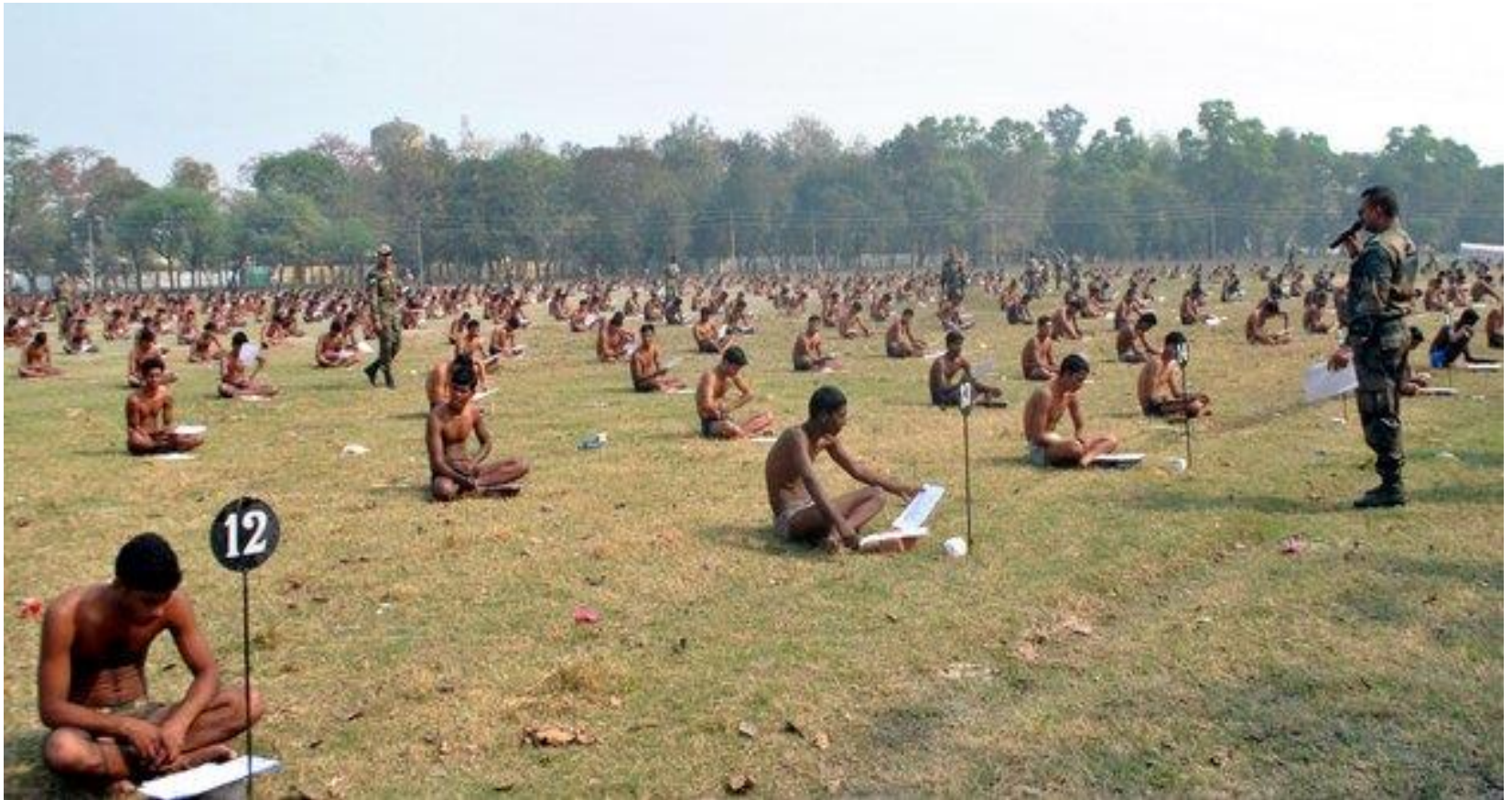
- +++ testing more frequently
- ++ testing with stakes
- + testing with feedback

# Effect of scale on testing benefits

- size of study population
  - small +0.34 over large
- scale of test administration
  - small-scale +0.14 over large-scale
- responsible level of government
  - local tests +0.29 over state tests



# Large-scale test, tight security



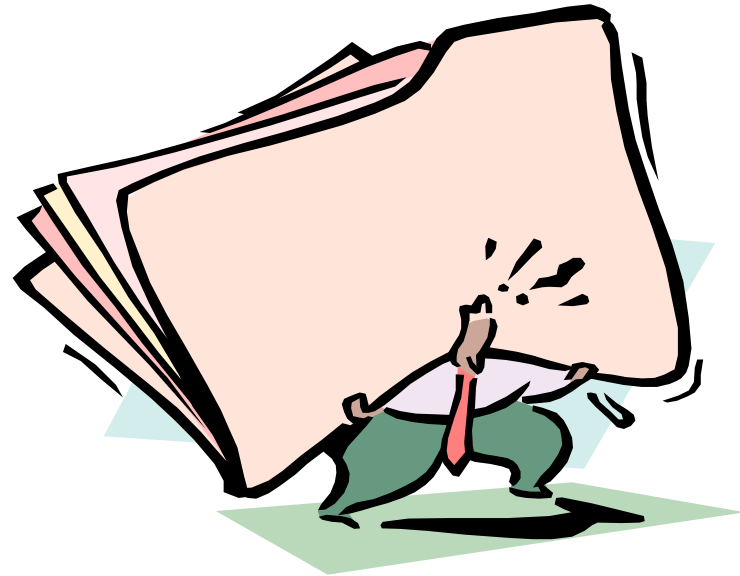
# Large-scale test, lax security



# Besides, systemwide tests are needed for other purposes, such as...

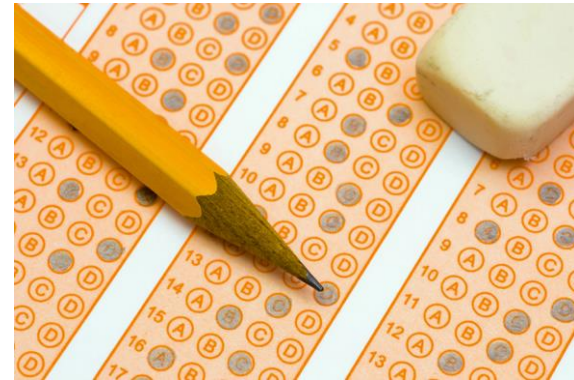
- ...selection to programs with limited number of places
- ...monitoring and system diagnosis
- ...workforce planning
- ...accountability
- ...credentialing

That's enough!





# Some large-scale test advantages



On per-student basis, inexpensive

Cognitive laboratory pre-testing possible

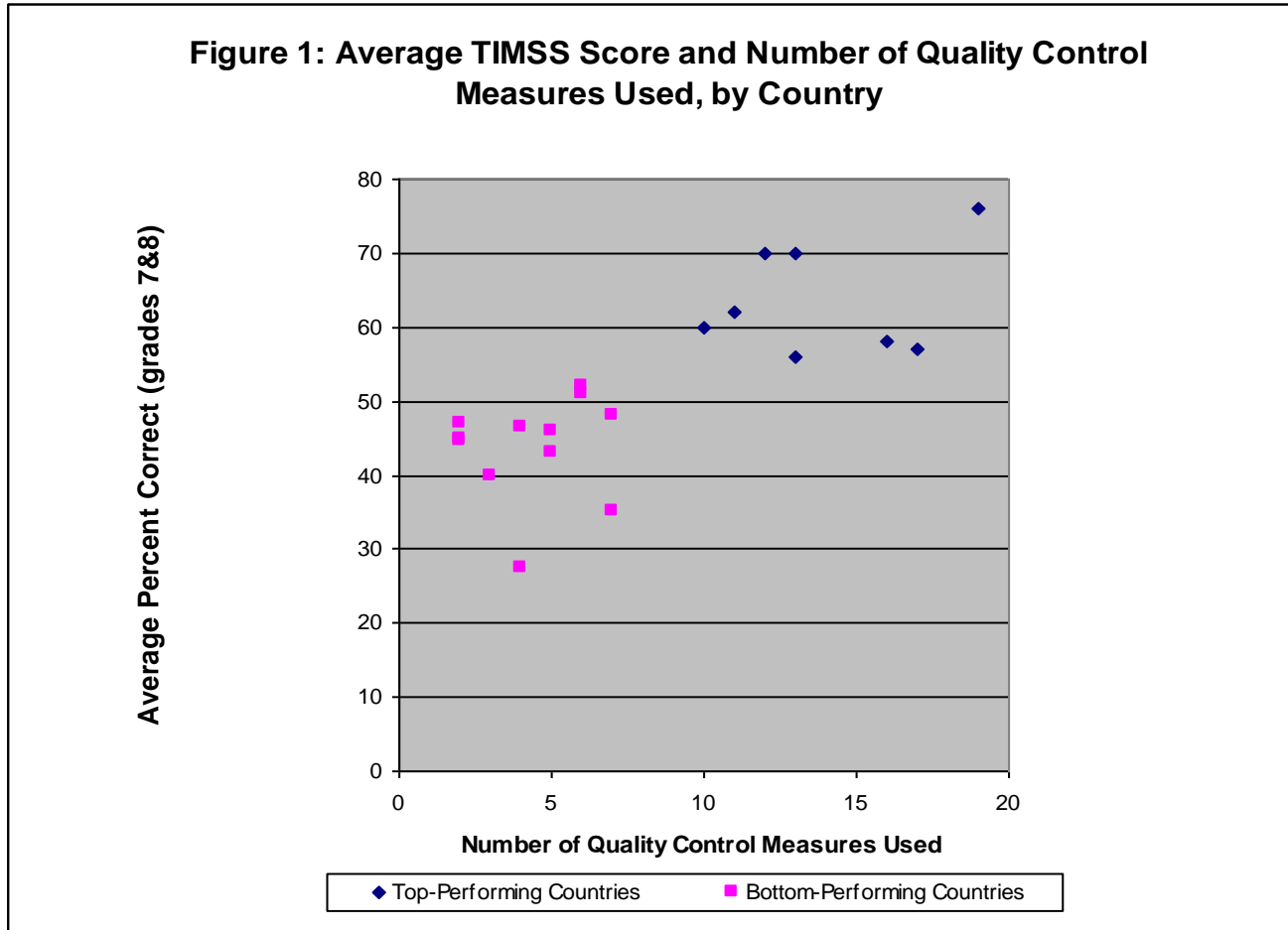
Standardization offers comparisons across schools and regions.

May produce high-quality items that schools and teachers can use.

**MOST IMPORTANT:**

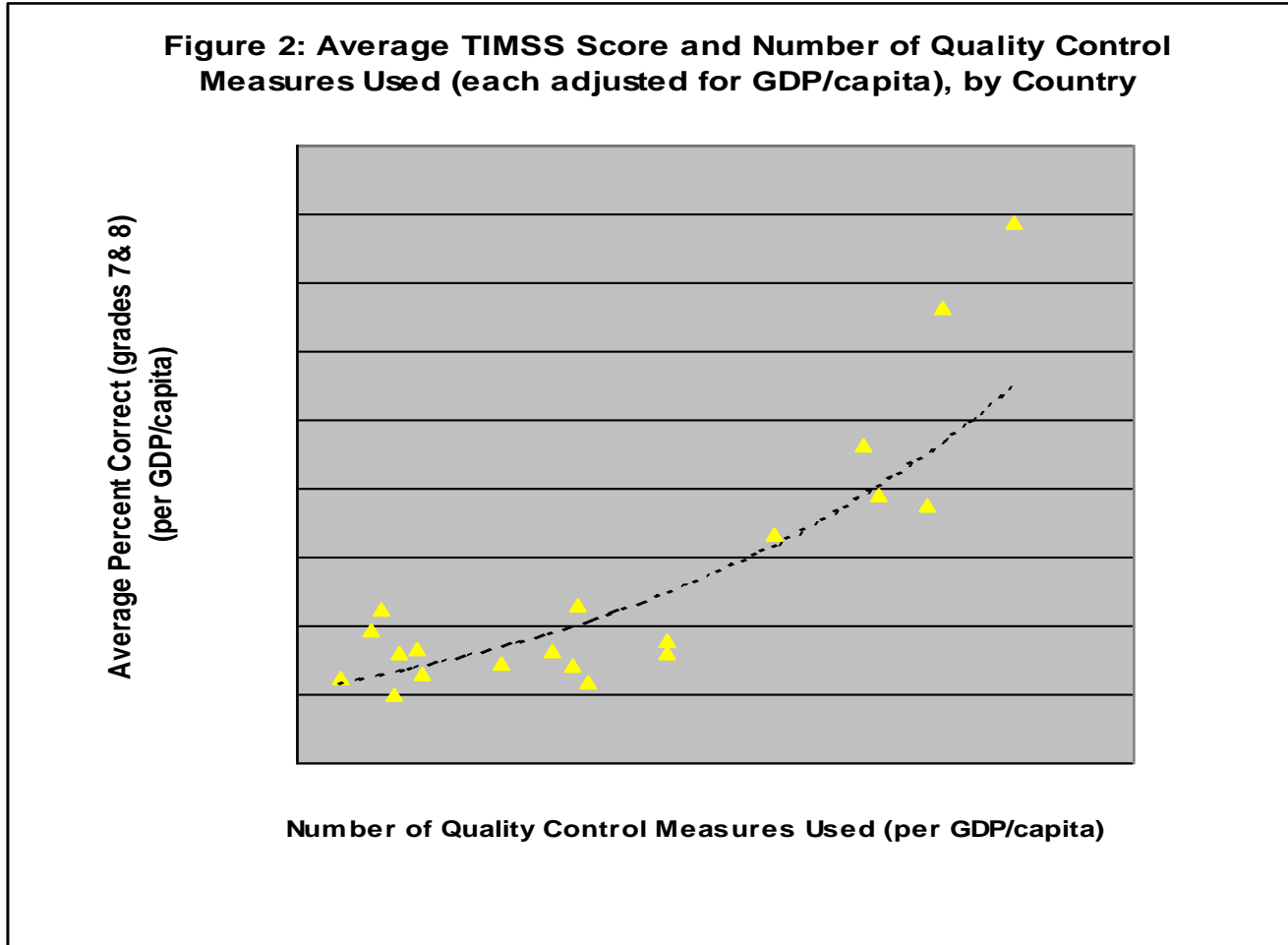
provides reliable, comparative information to all those not involved in a particular school

# The more systemwide decision points, the better ?



SOURCE: Phelps, Benchmarking to the best in mathematics, *Evaluation Review*, 2001

# Quality control has proportionally greater effect in poorer countries



SOURCE: Phelps, Benchmarking to the best in mathematics, *Evaluation Review*, 2001

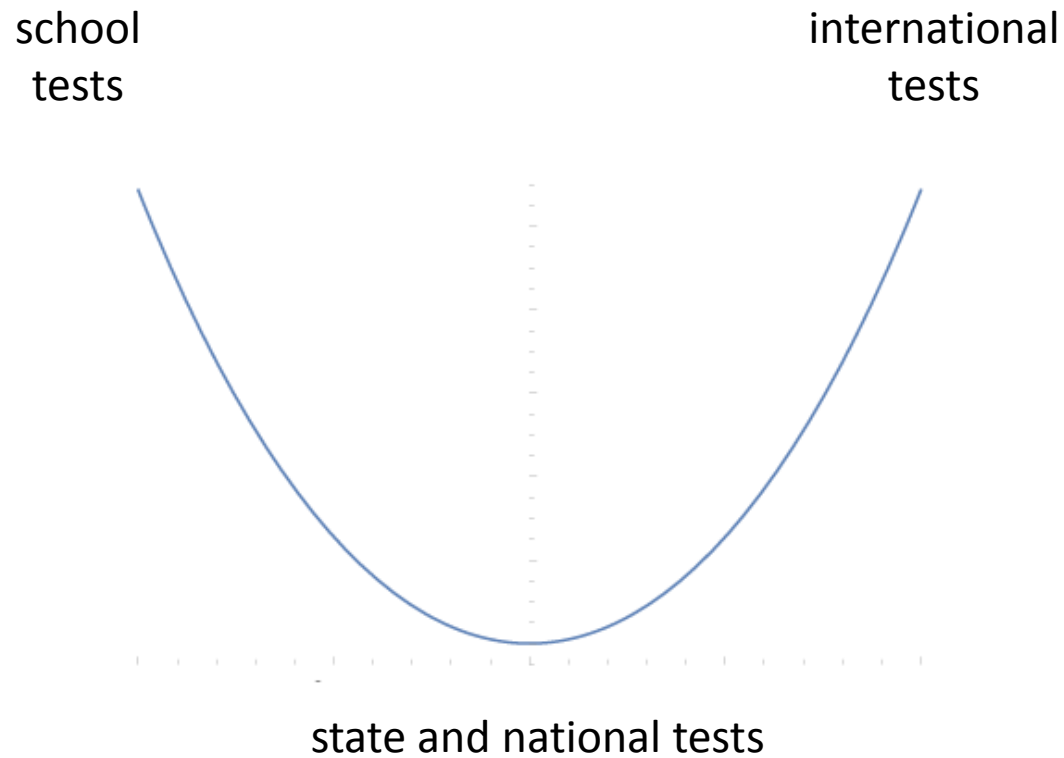
**IEA:** TIMSS, PIRLS, CIVED, SITES, ICILS,  
PPP, ECES, TEDS

**OECD PISA:** PISA, PISA for schools  
PISA for development

**World Bank:** READ, SABER  
...provides funding for PISA

# The effect of international testing programs

Freedom to design your testing



# OECD and World Bank are run by economists

How well do economists understand PSYCHO-metrics?

Some interesting examples:

Chile's national testing program, funded by the World Bank

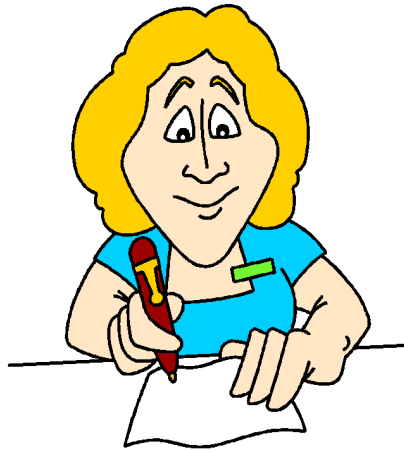
OECD's "Synergies for Better Learning" project



## Some interesting oddities:



# Designing an Assessment System



richard {at} nonpartisaneducation {dot} org