Using Simple Computer Simulations to Address Complex Assessment Problems

Stephen D. Salyards

Overview

- What is Statistical Simulation?
- Why Use Simulation?
- Brief History
- Basic Steps
- Assessment Applications

What is Statistical Simulation?

- Statistical simulation is based on the concept of “resampling”
- Resampling refers to the use of observed data, or of a data generating mechanism (such as a coin or die), to produce new hypothetical samples, the results of which can then be analyzed (Simon, 1999)
- Variations on the resampling theme:
  - Computer-intensive methods
  - Monte Carlo simulation
  - Bootstrap procedure
  - Permutation/randomization test
  - Exact Probability Test

Why Use Simulation?

- Simulation tends to be more robust and general than conventional techniques based on idealized, theoretical models
  - More flexible – can handle any problem conventional methods can handle – the reverse is not true
  - Normal-theory methods can be surprisingly inaccurate
- Simulation tends to be more transparent and requires fewer technical concepts and assumptions
  - Assumptions of conventional formulas are often hidden under a deep layer of mathematical theory
  - Simulation is now <u>the</u> benchmark by which we judge the performance of conventional procedures

History of Statistical Simulation

- Gosset (pseudonym “Student”, 1908) developed empirical probability distributions by resampling hundreds of times from a deck of shuffled cards containing a given dataset

- Gosset conducted his simulation research to develop reference distributions for cases where the “normal curve” was inappropriate

- A reference (or sampling) distribution is based on repeated random samples of size *n* and describes what values of a statistic will occur and how often

- A sampling distribution can be derived using probability theory (traditional approach) or by resampling actual or hypothetical data

History of Statistical Simulation (contd.)

- The great mathematician Ronald Fisher spent 7 years deriving theoretical formulas to approximate Gosset's empirical distributions (Student's *t*-test)

- We are no longer limited to using Fisher's formulas or the theoretical assumptions required to apply them

- Gosset's pioneering card-shuffling approach is back, only computers now do in seconds what once took months or years (Bennett, 1999)

- Key question: How often does your observed result occur as a matter of random sampling fluctuation?

Basic Steps

- Specify relevant universe (simulated population or process)
- Specify sampling procedure
  - Sample size
  - Number of samples
  - With or without replacement
- Compute statistic or descriptor of interest
- Resample, compute, and store results over several trials
- After completion, summarize the results in a histogram or probability distribution

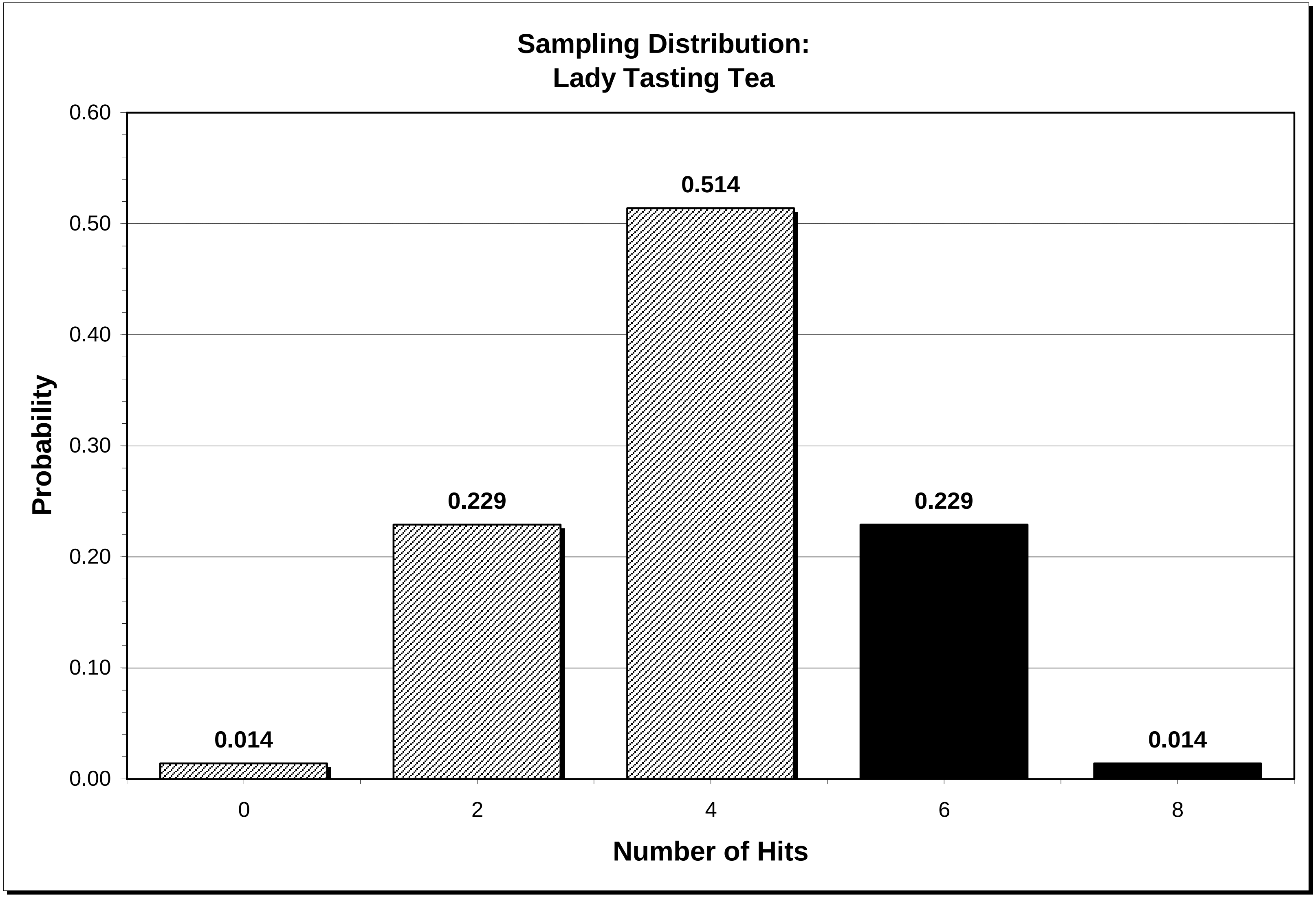
Basic Steps: Fisher's Tea Taster

- A woman attending a tea party claimed tea poured into milk did not taste the same as milk poured into tea
- Fisher set up an experiment to "test the proposition" (Salsburg, 2002)
- Eight cups of tea were prepared (four with tea poured first, and four with milk poured first) and presented randomly
- What is the probability of getting 6 correct guesses (hits) by chance alone?
- Design a simulation using a deck of eight cards (4 labeled milk-first, 4 labeled tea-first) or write a simple computer program

Basic Steps: Fisher’s Tea Taster (cont’d)

URN (0 0 0 0 1 1 1 1) actual	‘Tea cups; 0 = milk first, 1 = tea first
URN (0 0 0 0 1 1 1 1) guess	‘Guesses; 0 = milk first, 1 = tea first
REPEAT 1000	‘Repeat 1,000 times
SHUFFLE guess guess$	‘Shuffle the guesses
SUBTRACT actual guess$ diff	‘Check for matches, store result in diff
COUNT diff = 0 match	‘Zero indicates correct guess
SCORE match hit	‘Store number of hits for that trial
END	‘Stop after 1,000 trials
HISTOGRAM hit	‘X-axis shows number of hits

Basic Steps: Fisher's Tea Taster (cont'd)



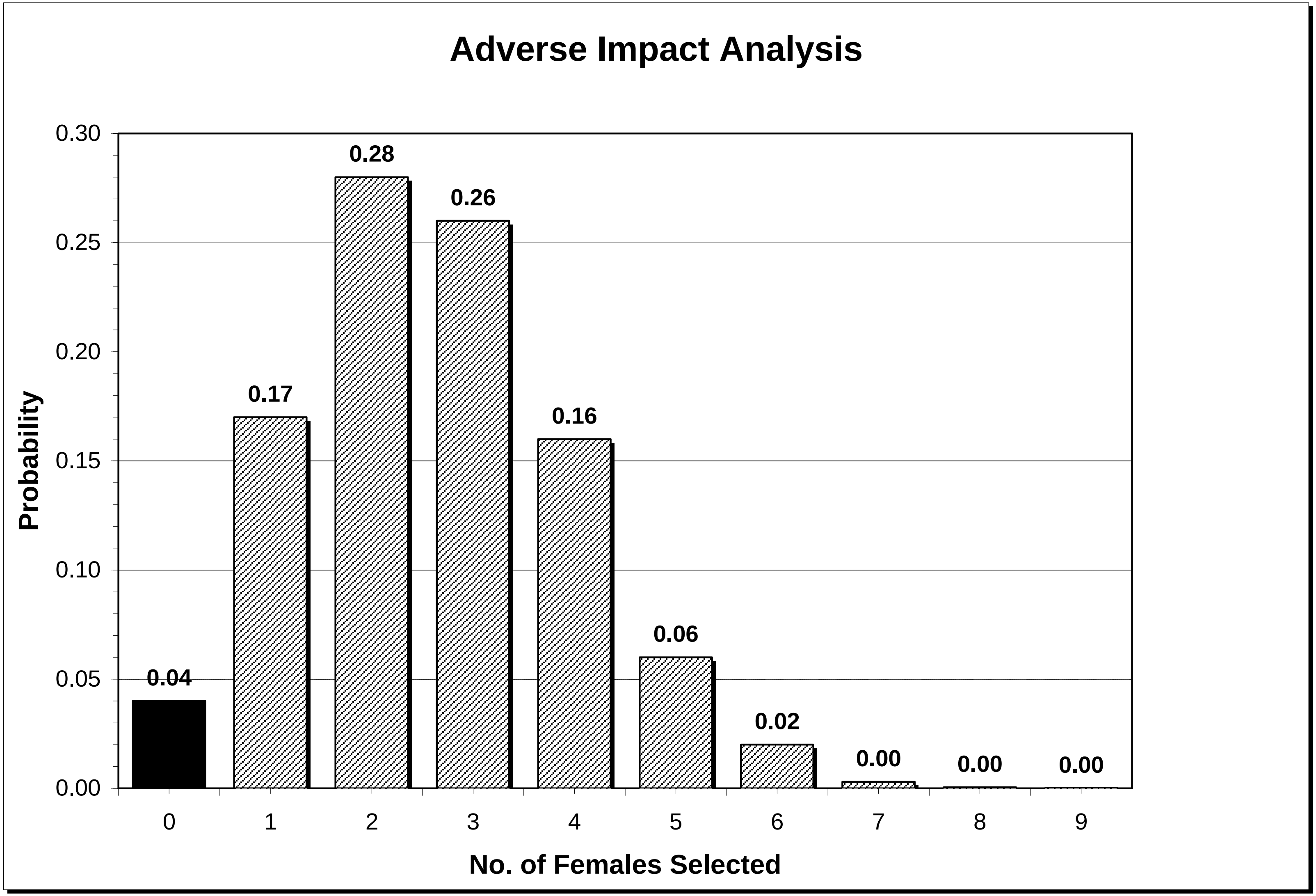
Assessment Applications

1. Adverse Impact (Single Applicant Pool)
2. Guessing on Matching Tests
3. Detection of Test Cheating
4. Score Categorization and Validity
5. Scale Compression and Information Loss
6. Sampling Distributions for New Statistics
7. Adverse Impact (Multiple Applicant Pools)

Application 1: Adverse Impact Analysis

- The four-fifths (or 80%) rule and the chi-square test for detecting adverse impact can disagree 10-40% of the time depending on sample size (York, 2002)
  - With small sample sizes, chi-square test has low power to detect differences in selection rates
  - With large sample sizes, four-fifths rule often fails to detect adverse impact
- Scenario: 80 men and 20 women apply for jobs, 13 applicants are selected
- What is the probability of <u>no</u> women being selected?
- Adverse impact? Four-fifths rule indicates “Yes”; Chi-square test says “No”

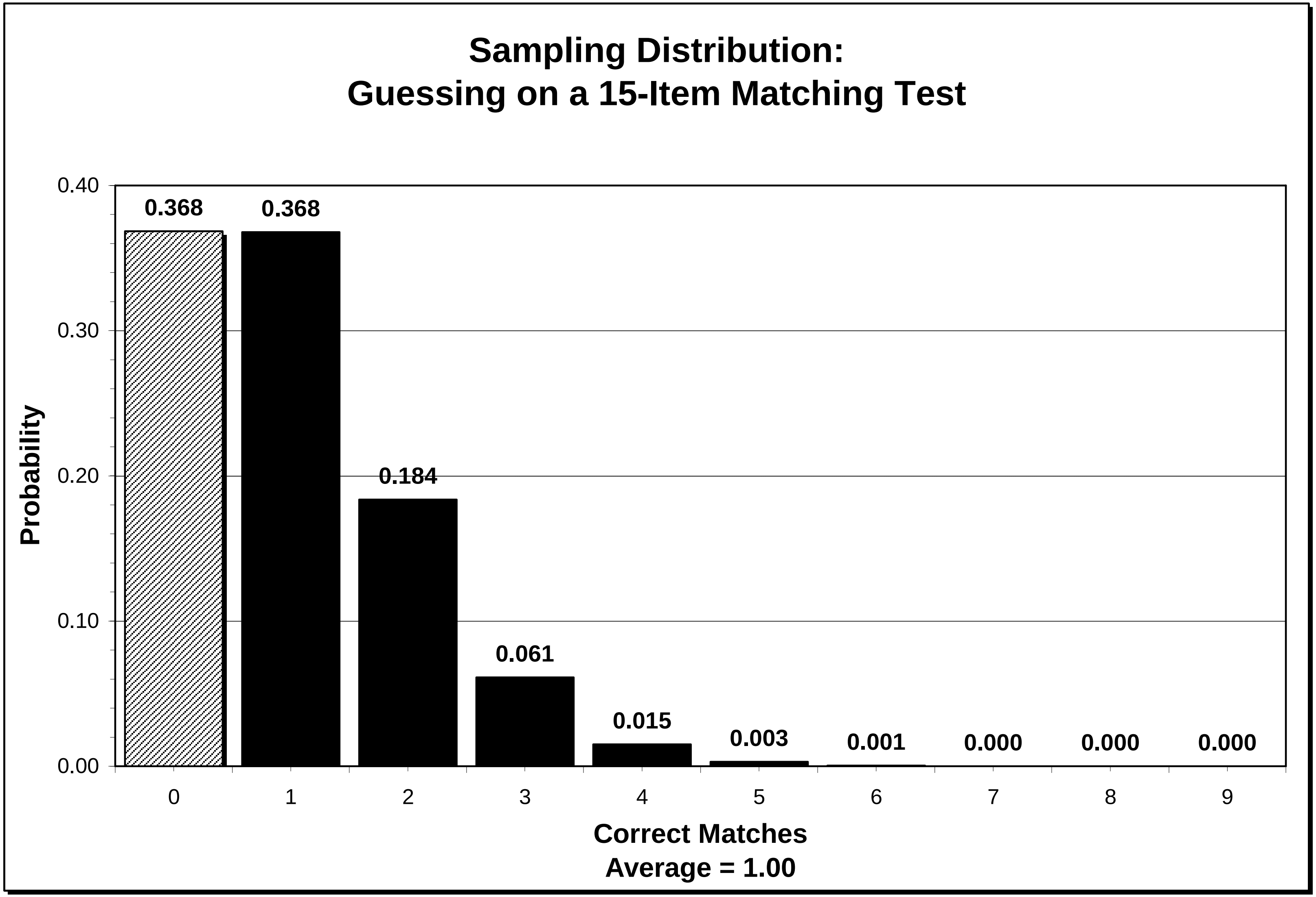
Application 1: Adverse Impact (cont'd)



Application 2: Matching Tests

- According to Haladyna (2004), nearly all measurement textbooks recommend using the matching item format
- For example, applicants are asked to match dates with historical events, parts with functions, terms with definitions
- There is surprisingly little research on this item format
  - Resilient to guessing?
  - What if the number of item stems and choices are unequal?
- Scenario: On a 15-item matching test, how many items can an applicant match correctly by chance alone?

Application 2: Matching Tests (cont'd)



Application 2: Matching Tests (cont'd)

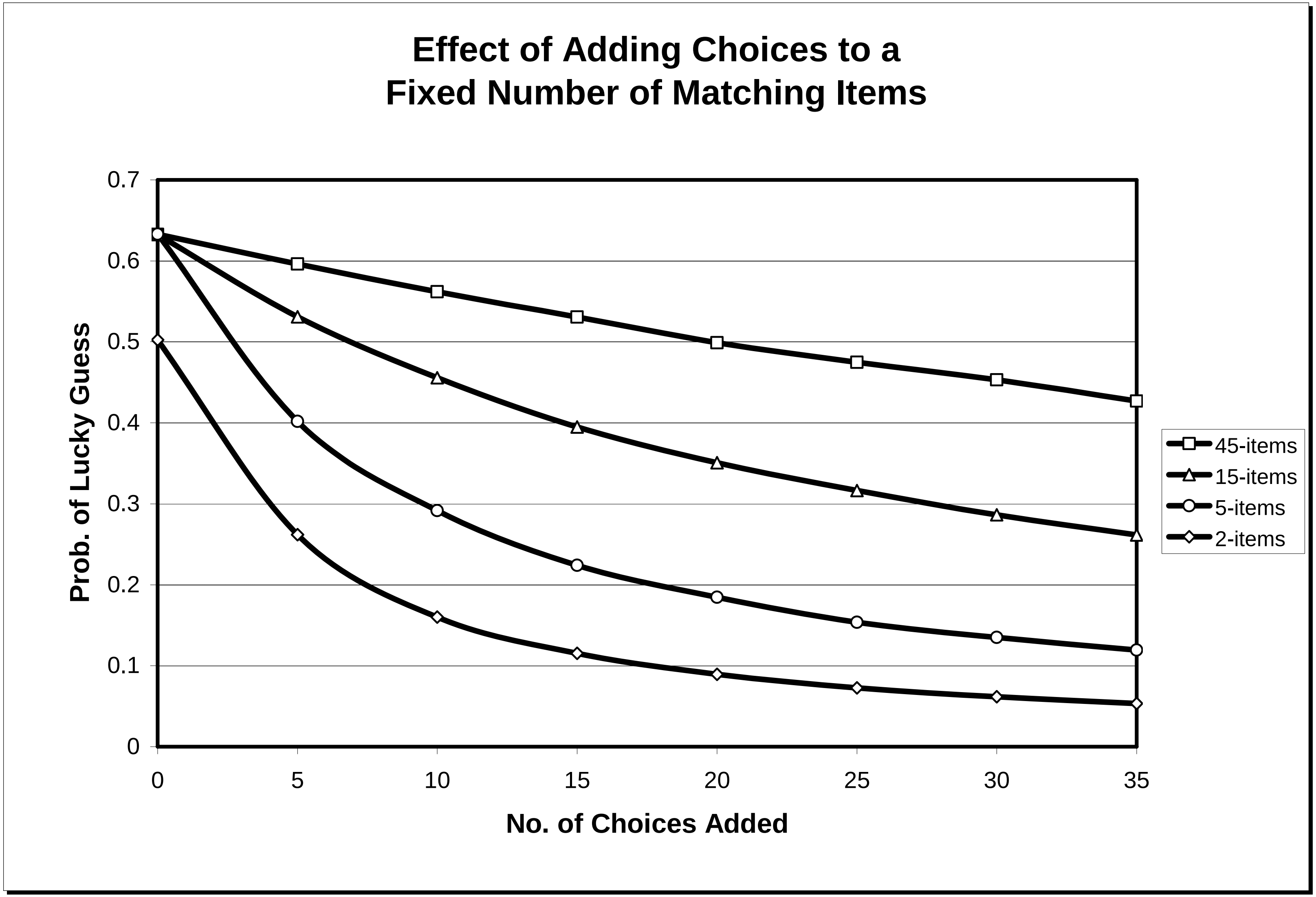
- Item writing guides often recommend making the number of options different from the number of item stems

- What is the expected effect on guessing from adding distractors (i.e., bogus options)?

- Is it worth the trouble to have item writers add plausible distractors to the list of correct options?

- Does the effect on guessing depend on the number of item stems?

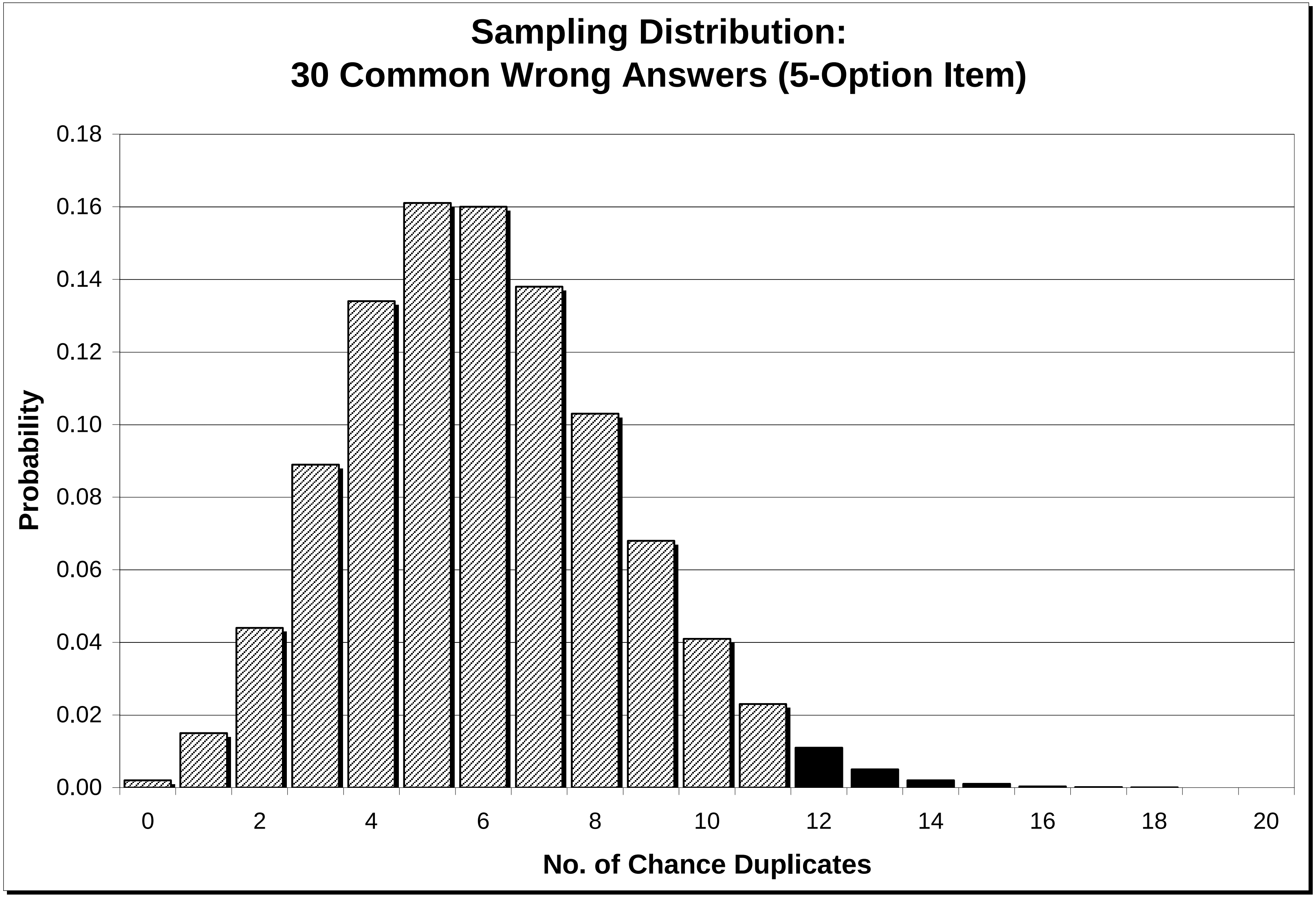
Application 2: Matching Tests (cont'd)



Application 3: Test Cheating

- How do you evaluate a claim by a test administrator that an applicant has copied answers from another?
- Some researchers have proposed looking at the similarity of incorrect responses (Bellezza & Bellezza, 1989)
- Distractors (wrong answers) are designed to seem equally plausible to those attempting to guess the right answer
- Applicants working independently (i.e., not copying from each other) do not tend to select the same distractors
- How many duplicate wrong answers would be expected by chance alone?

Application 3: Test Cheating (cont'd)



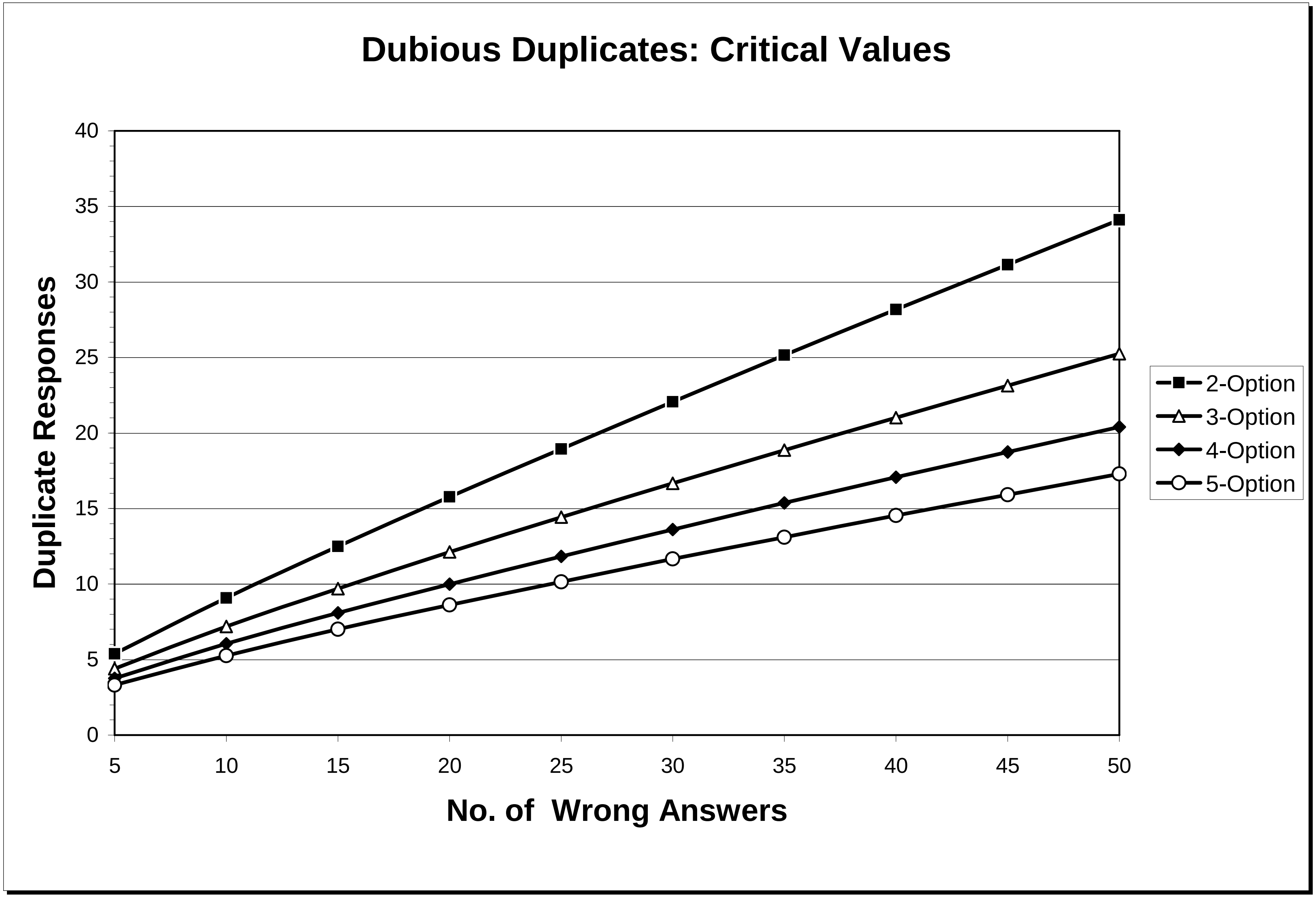
Application 3: Test Cheating (cont'd)

- The sampling distribution helps identify outliers (i.e., error patterns so similar that duplicates may have occurred through copying)

- We can set a threshold (i.e., critical value) where the number of duplicates is so extreme that it is unlikely to have occurred by chance (e.g., only one chance in a hundred)

- Does the number of multiple choice options affect the number of expected duplicates?

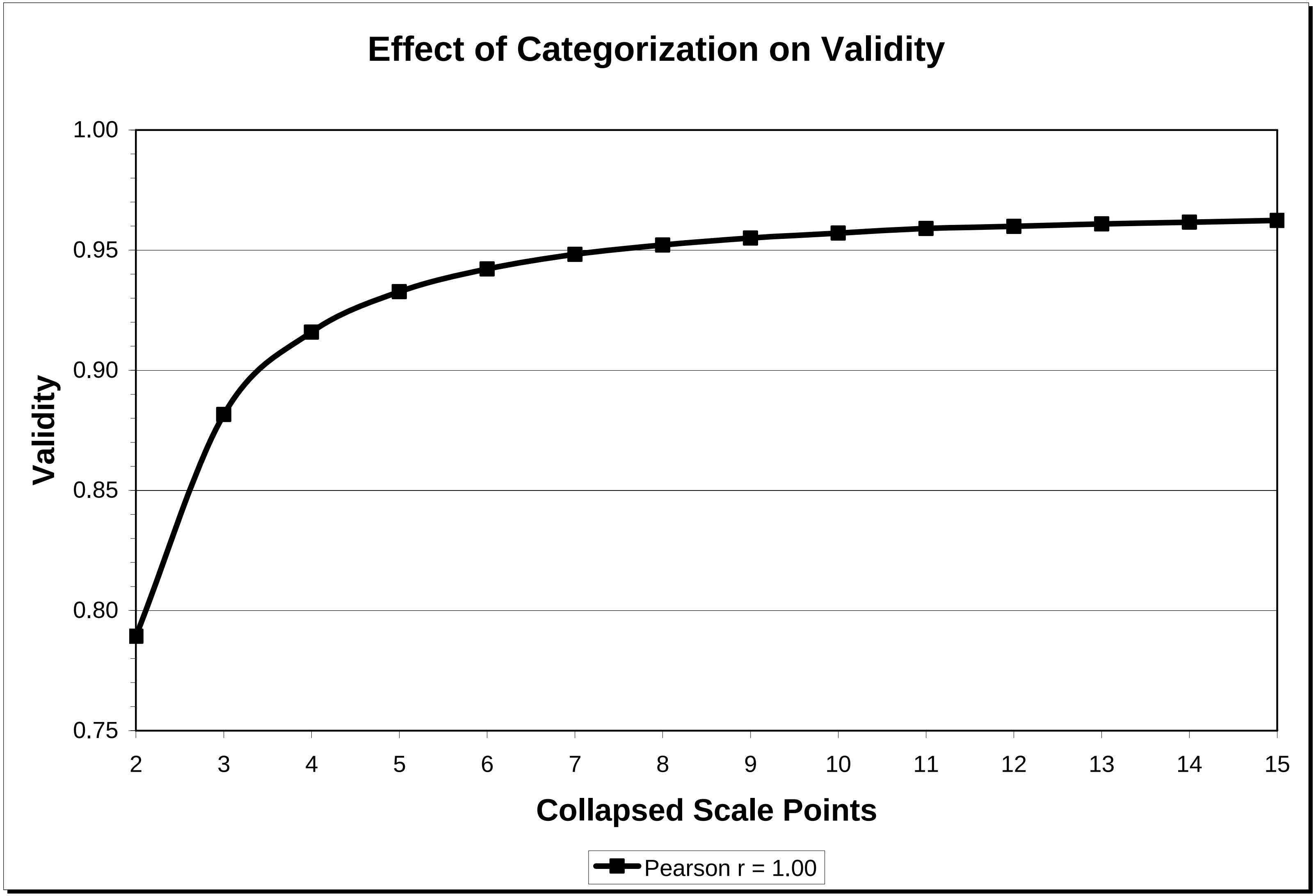
Application 3: Test Cheating (cont'd)



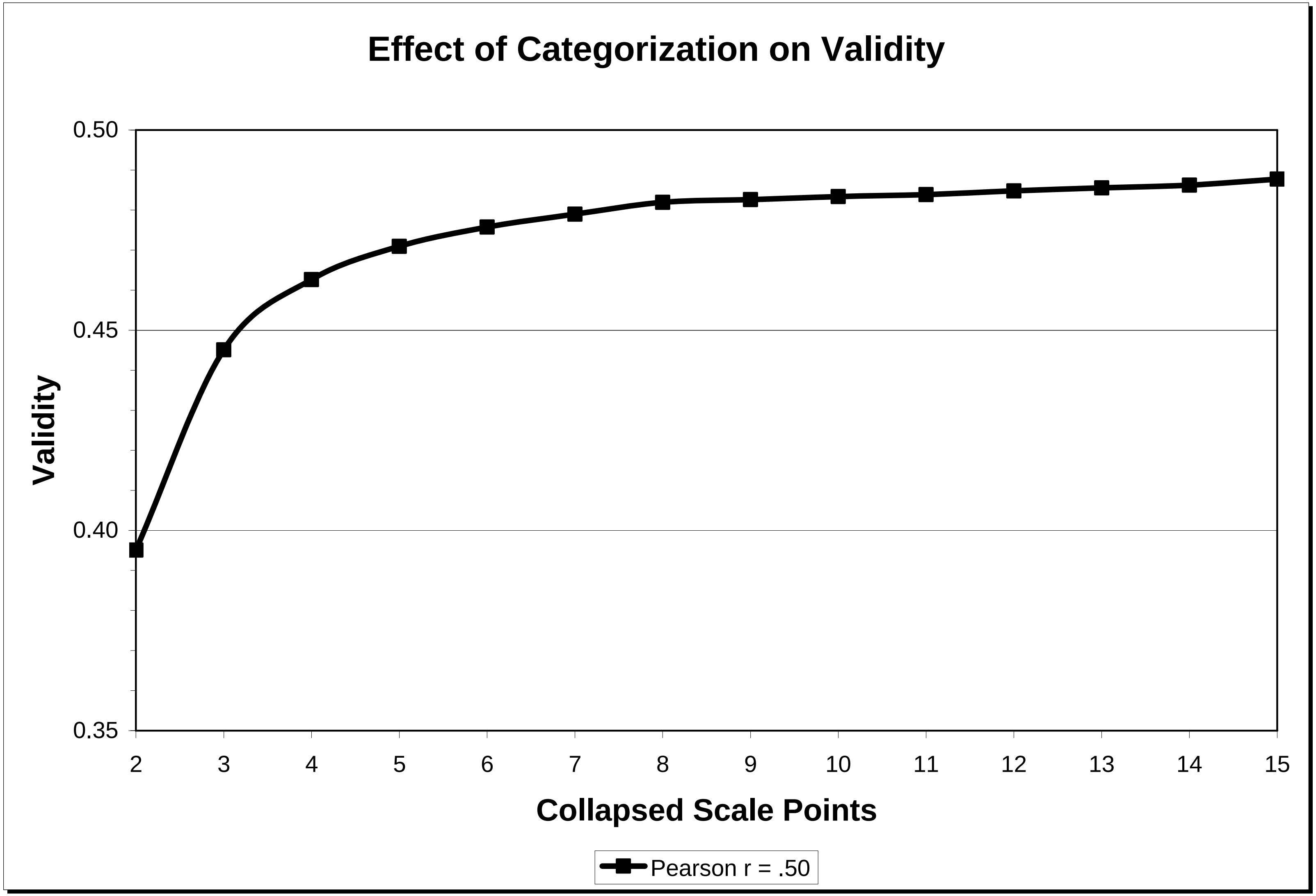
Application 4: Categorization and Validity

- Score banding involves collapsing a continuous distribution of scores into discrete categories (e.g., High, Medium, Low)
- How much information loss can be expected from categorizing continuous test scores?
- Scenario: Take a 100-point scale, collapse into categories, and then correlate it with its categorized self
- Any loss of information should cause the resulting correlation to differ from 1.00 (i.e., a perfect correlation)
- The magnitude of the difference provides an index of information loss

Application 4: Categorization and Validity (cont'd)



Application 4: Categorization and Validity (cont'd)



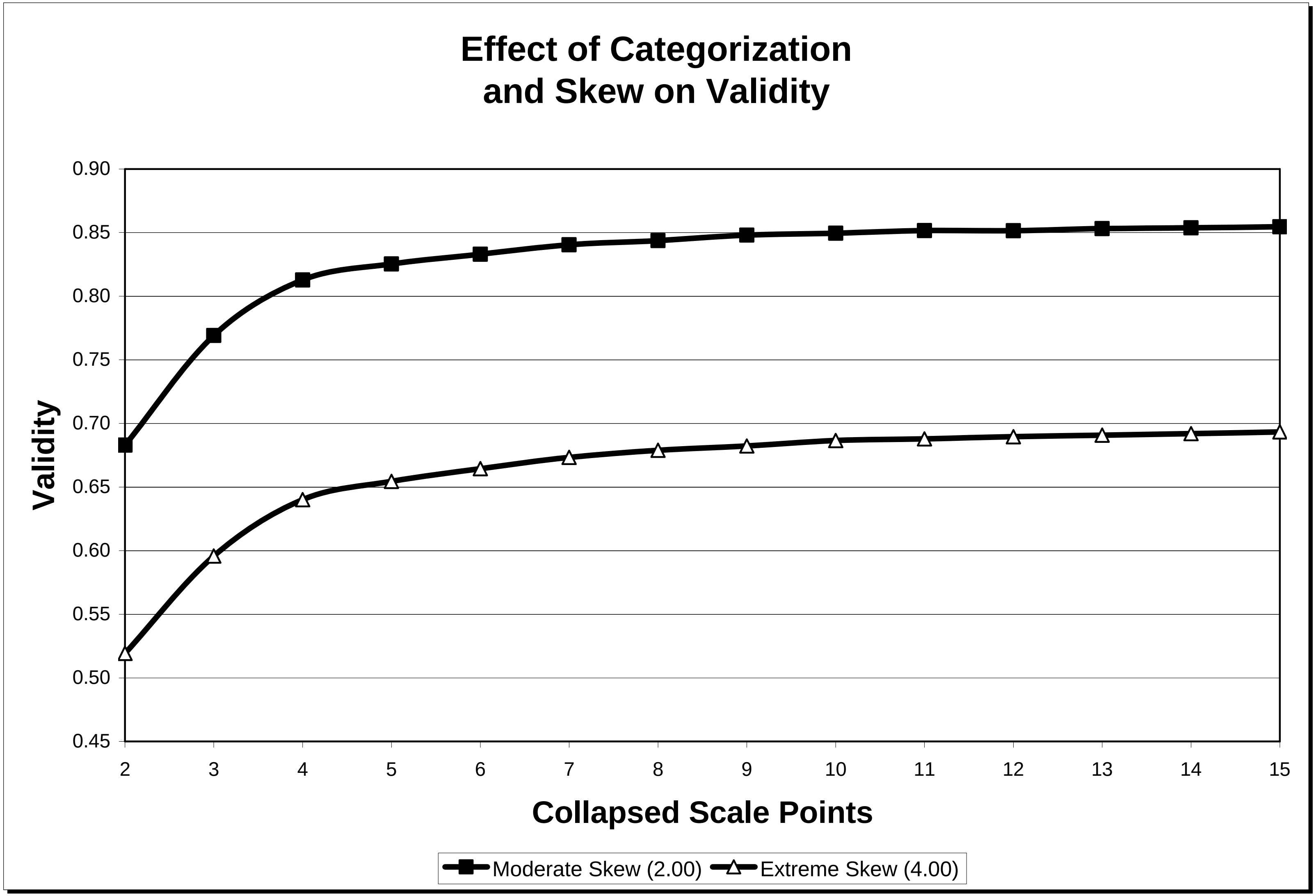
Application 4: Categorization and Validity (cont'd)

- Performance ratings are often highly skewed in accordance with the Lake Wobegon Effect (e.g., "All of my employees are above average")

- What happens to a skewed measure that is then categorized?

- Simulation allows us to quantify the impact of scaling in the presence of many other factors (e.g., measurement error, nonnormal distributions) for any statistic (e.g., Cronbach's alpha, variance, rWG)

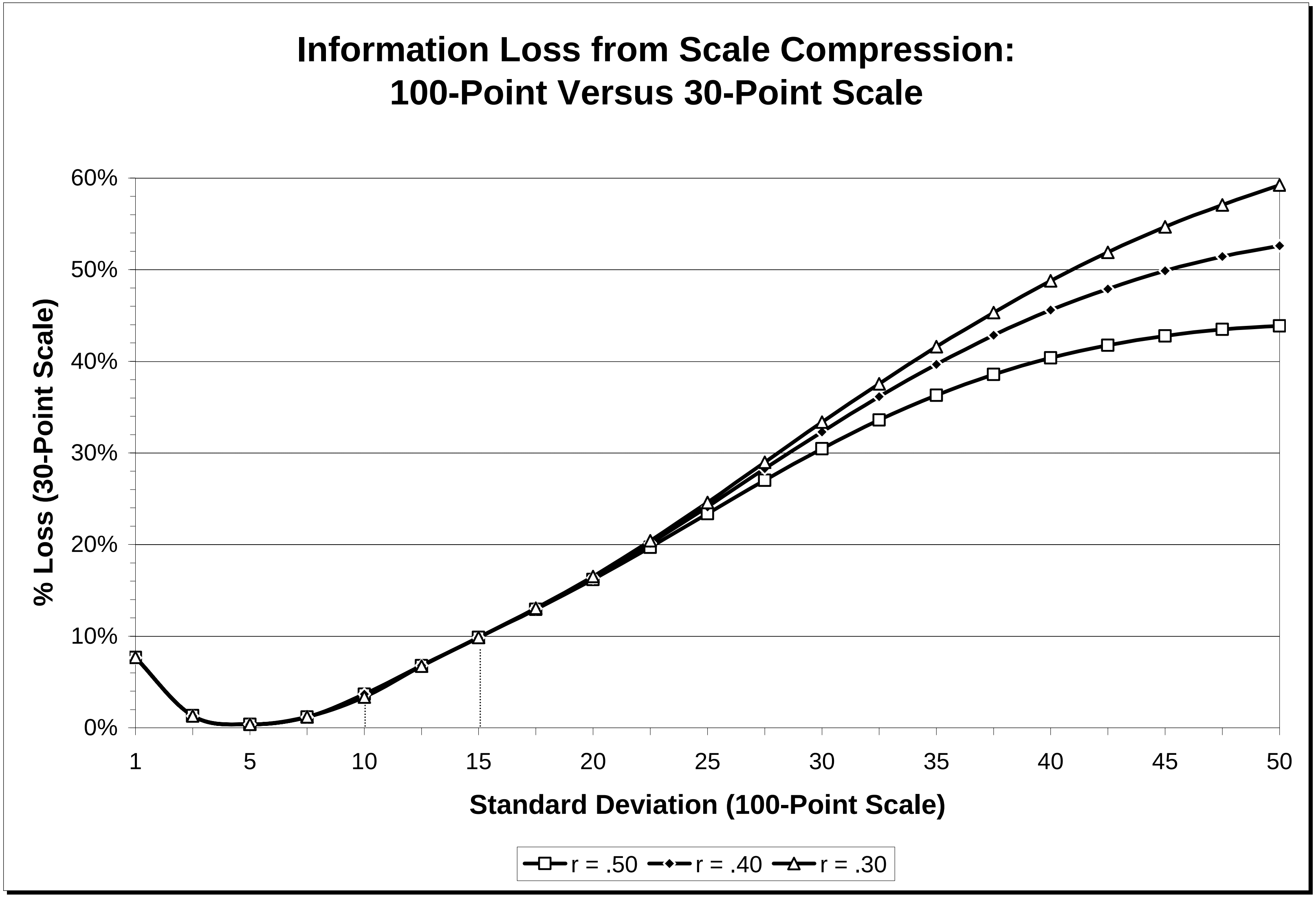
Application 4: Categorization and Validity (cont'd)



Application 5: Scale Compression

- According to the Partnership for Public Service (2004), many agencies are awarding “70 points, out of 100, to candidates simply for satisfying minimum qualifications”
- This practice only leaves 30 points for further assessments that rate and rank applicants
- PPS argues that “this kind of compression significantly erodes the power of any assessment tool to make meaningful distinctions in likely candidate performance”
- How much information loss results from this type of compression?

Application 5: Scale Compression (cont'd)



Application 5: Scale Compression (cont'd)

- The simulation assessed information loss for score distributions with varying levels of spread and predictive validity
- Validity of the original 100-point scale had little impact on the amount of information loss when collapsing to a 30-point scale
- For realistic levels of score variation among applicants (typical SDs run between 10 and 15), little information is lost due to scale compression (less than 10%)
- Information loss may or may not be significant, depending on the amount of variation observed in the original scores

Application 6: New Statistics

- Simulation can be used to derive the distributional properties of *any* statistic, even new or home-grown statistics with no known sampling distribution

- The sampling distribution for some statistics cannot be derived mathematically (e.g., the median) or can only be crudely approximated using normal-theory assumptions

- What does a practitioner do when the assumptions of the statistical test are violated (e.g., skewed data, unequal variances)?

- Simulation gives us a way to solve these problems

Application 6: New Statistics (cont'd)

- *rWG* is the most widely used measure of interrater agreement for Likert-type scales (Kline, 2005)

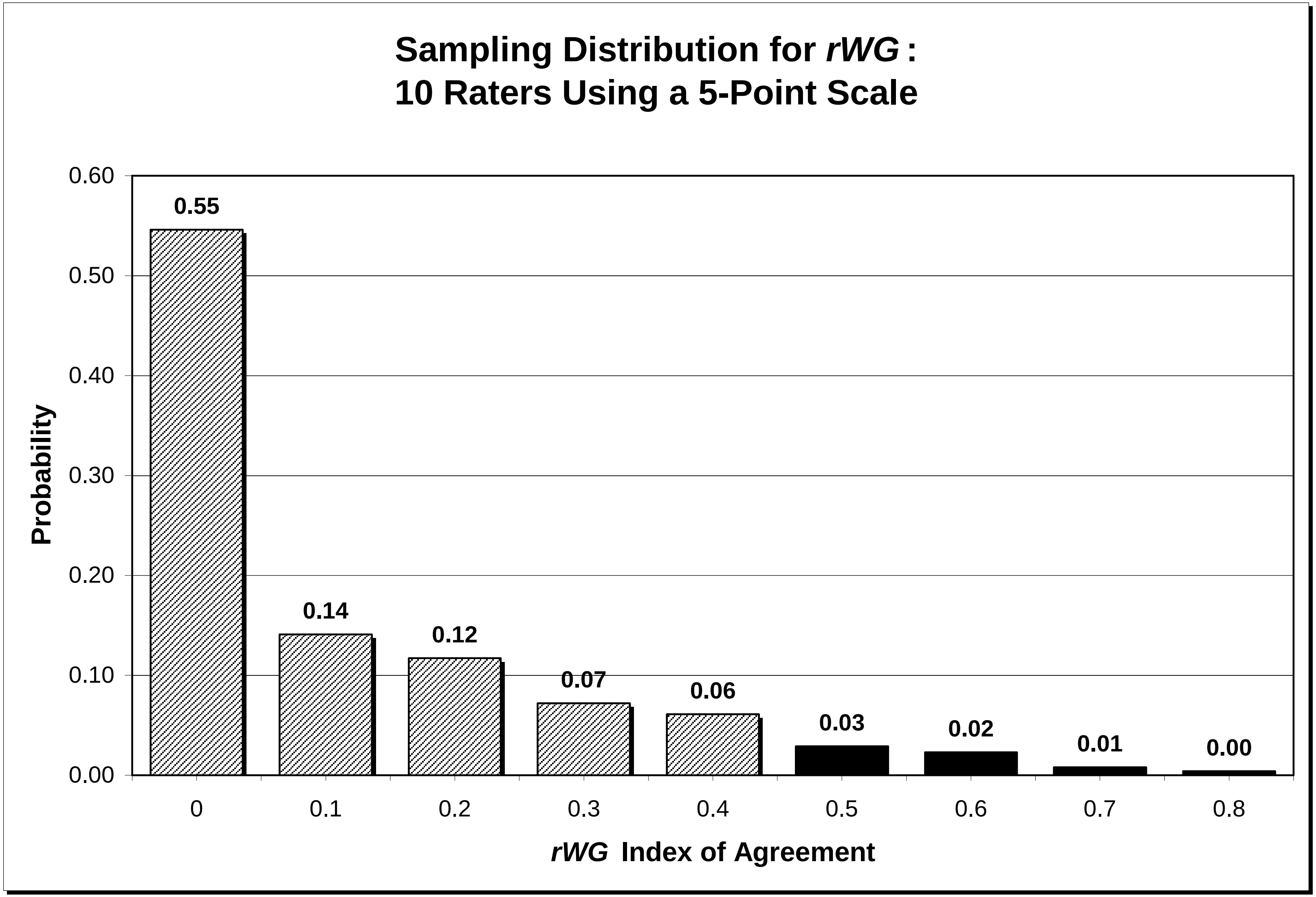
- *rWG* compares the variability in observed ratings to the expected variability of randomly generated ratings:
  - *rWG* = 1 – [Var (observed) / Var (random)]

- Unfortunately, there is no consensus on a statistical significance test for *rWG* (Dunlap et al., 2003)

- Scenario: A panel of 10 job experts is asked to rate the content validity of test items using a 5-point Likert-type scale

- What value of *rWG* must be achieved to reach statistical significance?

Application 6: New Statistics (cont'd)



Application 7: Adverse Impact and Multiple Events

- In a given year, employers are often faced with multiple selection events for the same job
- Summing data across these events treats each applicant as if he or she competed in each selection event (Siskin & Trippi, 2005)
- Summing treats selections as if they were made from a *single* pool of applicants rather than from multiple pools and can produce biased, misleading results
- According to Gilmartin & Claudy (1985), “the single pool approach is inherently wrong” and selection probabilities will be incorrect

Application 7: Adverse Impact (cont'd)

- The aggregation method gaining acceptance by the courts (see Gilmartin, 1991) involves combining the *sampling distributions* from each selection event

- Exact probabilities can be obtained using either additive convolution techniques or computer simulation (Poe et al., 2005)

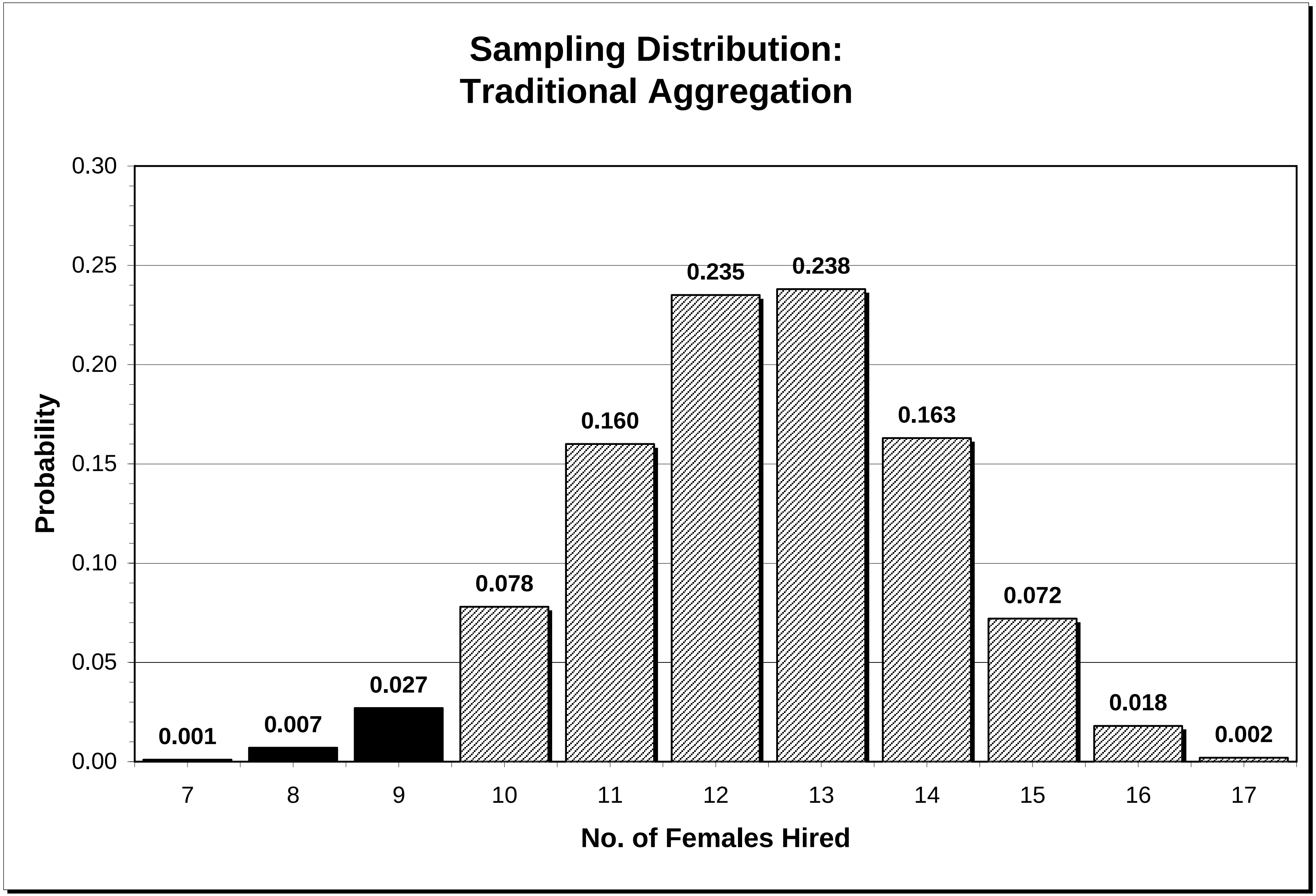
Application 7: Adverse Impact (cont’d)

Table 1
Aggregation of Selection Data: Traditional versus Multiple Exact Method

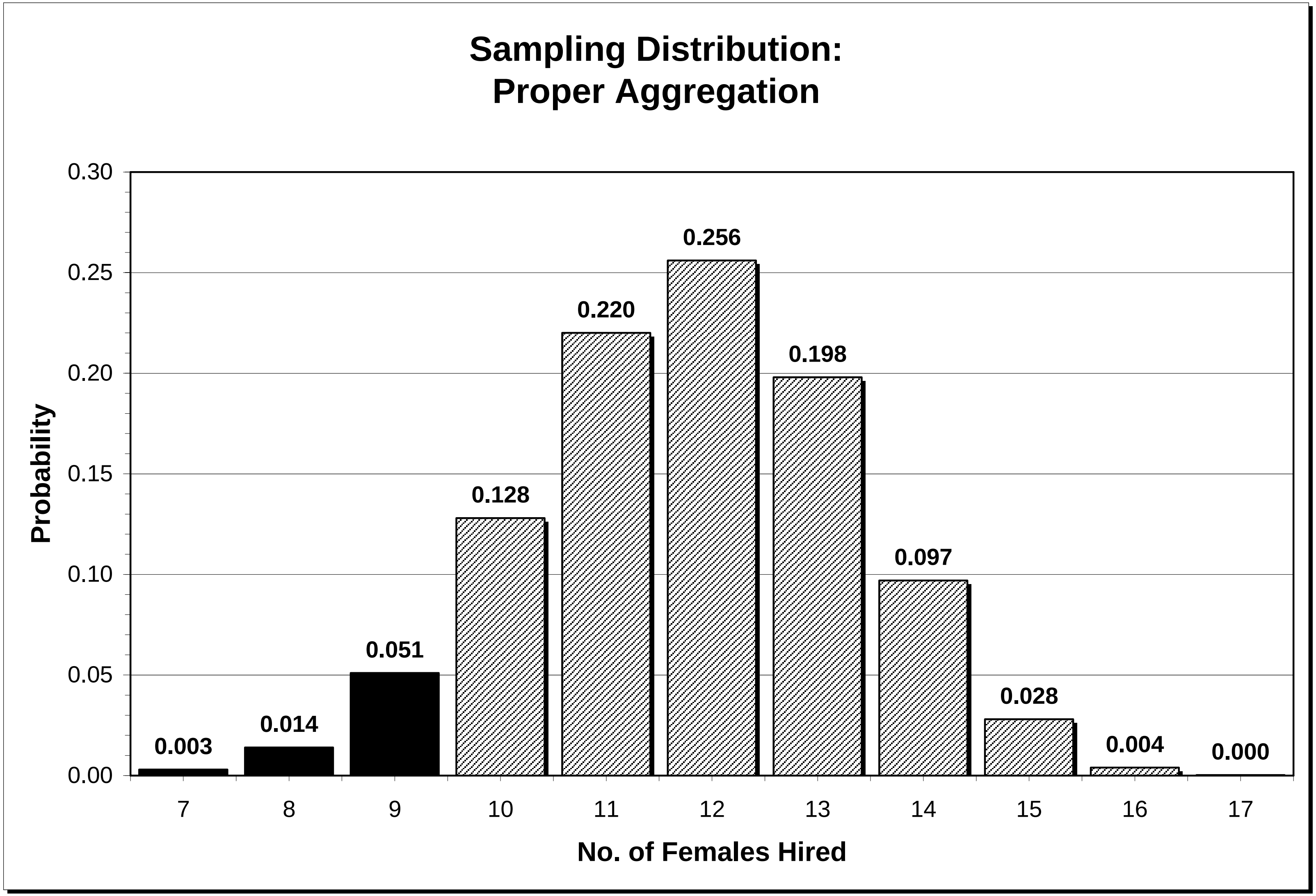
	Applicant Pool		Hires		Expected		Adverse Impact?		
Selection Event	**Females**	**Males**	**Females**	**Males**	**Females**	**Shortfall**	**4/5ths**	**Sig.**	***p***
Jan/2004	6	25	5	23	5.42[a]	.42	No	No	.49
Aug/2004	5	15	1	11	3.00[b]	2.00	Yes	No	.06
Nov/2004	6	10	3	6	3.38[c]	.38	No	No	.55
Overall									
Multiple Exact	-	-	9	40	11.80	2.80	-	No	.07
Traditional	**17**	**50**	**9**	**40**	**12.43**	**3.43**	**Yes**	**Yes**	**.03**

[a] [6 / (6 + 25)] x 28
[b] [5 / (5 + 15)] x 12
[c] [6 / (6 + 10)] x 9

Application 7: Adverse Impact (cont'd)



Application 7: Adverse Impact (cont'd)



Summary

- Simulation methodology can be used to address a number of practical assessment problems that are too complex, too time-consuming, or even impossible to answer using traditional analytical methods

- Traditional methods can only be trusted under certain, restricted circumstances, whereas simulation is subject to far fewer assumptions and constraints

- Resampling approaches have gained wide acceptance by statisticians and are being introduced in an increasing number of textbooks (e.g., Howell, 2002; Lunneborg, 2000)

References

Bellezza, F. S. & Bellezza, S. F. (1989). Detection of cheating on multiple-choice tests by using error similarity analysis. *Teaching of Psychology*, *16*, 151-155.

Bennett, D. J. (1999). *Randomness*. Cambridge, MA: Harvard University Press.

Dunlap, W. P., Burke, M. J., & Smith-Crowe, K. (2003). Accurate tests of statistical significance for *rWG* and AD interrater agreement indices. *Journal of Applied Psychology*, *88*, 356-362.

Gilmartin, K. J. (1991). Identifying similarly situated employees in employment discrimination cases. *Jurimetrics Journal*, *31*, 429-440.

Gilmartin, K. J. & Claudy, J. G. (1985). PROC MULTIEVENT: Multiple events exact probability test validation report. Palo Alto, CA: American Institutes for Research.

Haladyna, T. H. (2004). *Developing and validating multiple-choice test items (3rd ed.)*. Mahwah, NJ: Lawrence Erlbaum.

Howell, D. C. (2002). *Statistical methods for psychology (5th ed.)*. Pacific Grove, CA: Duxbury.

Kline, T. (2005). *Psychological testing: A practical approach to design and evaluation*. Thousand Oaks, CA: Sage.

References (cont'd)

Lunneborg, C. E. (2000). *Data analysis by resampling: Concepts and applications*. Brooks/Cole: Pacific Grove, CA.

Partnership for Public Service (2004). *Asking the wrong questions: A look at how the Federal government assesses and selects its workforce.* Washington, DC: Author.

Poe, G. L., Giraud, K. L., & Loomis, J. B. (2005). Computational methods for measuring the difference of empirical distributions. *American Journal of Agricultural Economics*, *87*, 353-365.

Salsburg, D. S. (2002). *The lady tasting tea: How statistics revolutionized science in the twentieth century.* New York, NY: Owl Books.

Simon, J. L. (1999). *Resampling: The new statistics (2nd ed.).* Resampling Stats: Arlington, VA.

Siskin, B. R. & Trippi, J. (2005). Statistical issues in litigation. In F. J. Landy (Ed.), *Employment discrimination litigation: Behavioral, quantitative, and legal perspectives* (pp. 132-166). San Francisco, CA: Jossey-Bass.

Student. (1908). The probable error of a mean. *Biometrika*, *6*, 1-25.

York, K. M. (2002). Disparate results in adverse impact tests: The 4/5ths rule and the chi square test. *Public Personnel Management*, *32*, 253-262.