Performance characteristics of several “rules” for self-interpretation of proficiency testing (PT) data

R.N. Carey[1], G.S. Cembrowski[2], C.C. Garber[3], **Z.Zaki[4];**

[1]Peninsula Regional Medical Center, Salisbury, MD, [2]University of Alberta Hospital, Capital Health Authority, Edmonton, AB, CANADA, [3]Quest Diagnostics Teterboro, NJ, [4]East Carolina University School of Medicine, Greenville, NC.

ABSTRACT

Objective: To determine which rules provide optimal performance for interpreting data from PT with 5 samples per event even when a participant’s performance is acceptable according to the limits set by the PT organization.

Methods: We used Monte Carlo computer simulation techniques to study the performance of several rules for interpreting PT data, relating their error detection capabilities to (1) analytical quality of the method, (2) probability of failing PT, and (3) ratio of the group standard deviation to the average intralaboratory standard deviation. Analytical quality is indicated by the ratio of the intralaboratory standard deviation (s_i) to the PT allowable error (E_A). Failure of PT was defined (CLIA) as an event when 2 or more results out of a total of 5 exceeded acceptable limits. We simulated 10,000 participant PT events with and without bias and increased random error. Minitab statistical software (State College, PA) was used to simulate PT for potassium, creatine kinase, and iron. Average s_i/E_A values ranged from 0.14 (high quality, potassium) to 0.34 (marginal quality, iron). Average group standard deviation (s_g) to intralaboratory standard deviation (s_i) ratio values ranged from 1.3 to 2.7. We studied the effects of varying amounts of systematic and random error. We investigated "counting" rules based on standard deviation index (SDI) limits: 1_{2SDI}, $1_{2.25SDI}$, 1_{3SDI}, 2_{2SDI}, 1_{2SDI} in 2 PT events; range rules R_{3SDI}, R_{4SDI}, and mean rules $X_{1.0SDI}$ and $X_{1.5SDI}$. We also investigated two rules based on E_A: $1_{75\%\bullet EA}$ (one result or more has error exceeding 75% of E_A), and $5_x\&1_{50\%\bullet EA}$ (all results are on the same side of the mean and one or more has error exceeding 50% of E_A).

Results: For high quality methods, the probability of failing PT is nearly zero until errors are many times s_i, and traditional counting rules perform well for interpreting PT data to detect significant errors. As method quality is reduced, traditional counting rules lose power to detect errors. For marginal quality methods, the probability of failing PT is higher (approximately 30% when the analytical error is a shift of the mean by $2\bullet s_i$) than the probability that counting rules will detect significant errors (approximately 0% for the 1_{3SDI} or 2_{2SDI} rule for a shift of $2\bullet s_i$). The sensitivity of counting rules to error is reduced as s_g/s_i increases. We recommend screening PT data with the $1_{75\%\bullet EA}/R_{4SDI}/X_{1.5SDI}$ combination rule. If the PT data cause rejection by any of these three rules, other rules can be used to determine whether the error is random or systematic. False rejections are a problem when this combination rule is used with marginal quality methods (up to 18% for $s_i/E_A = 0.34$); however, the probability of detecting a shift of $2\bullet s_i$ is over 90%. False rejections are nearly zero for high quality methods. In real PT data, we found 14 combination rule rejections in 185 challenges, a rejection rate of 8%. On investigation, the majority of these rejections or “near misses” were consistent with assay problems, although only one result exceeded PT allowable error limits.

Introduction

Laboratories must achieve passing grades in proficiency testing (PT) to maintain licensure and accreditation; however, results of PT provide information beyond pass/fail. Passing PT data can be examined in more detail by the laboratory to detect analytical bias and imprecision.

We studied the performance of several algorithms ("PT rules," analogous to QC rules) for interpreting PT data when there are 5 samples per PT event.

We were interested in answering 3 questions:

1. When errors occur, what is the probability of failing PT?
2. What size of error do I need to be able to detect to prevent future PT failures?
3. What PT rules work best for detecting significant errors?

The answers depend on the ratio of the laboratory's internal standard deviation for a method, s_i, to the allowable error for the analyte defined by the PT provider, E_A. When s_i is small relative to E_A, the probability of failing PT is low, and the "quality" of the method is high.

Analytical Method Quality[1,2]

s_i / E_A	Method Quality
$\geq$0.50	Unacceptable
0.33 - 0.50	Marginal
0.25 - 0.33	Fair
0.17 - 0.25	Good
<0.17	Six Sigma

[1] Assumes no bias relative to peer group

[2] Adapted from Westgard

Methods

- We used Monte Carlo computer techniques to simulate PT events of 10,000 participants.

- For each event, we simulated the baseline conditions of no error, and then added random or systematic error. Minitab statistical software (State College, PA) was used to simulate PT results for potassium, CK, and iron.

- The baseline data were derived from a 1990 CAP PT survey and successive year’s QC program using the same materials on a popular multi-channel analyzer.

- These data enabled accurate estimates of a typical laboratory’s internal standard deviation, s_i, and the peer group’s overall standard deviation, s_g, with PT materials for each analyte studied.

Analytes Studied

Potassium, CK, and iron span the range of quality from Six Sigma (potassium) to marginal (iron).

Analyte	Units	E_A	Source of E_A	s_i (CV_i,%)	Mean s_i/E_A
K	mmol/L	0.5	CLIA	0.07 (1.6)	0.14
CK	U/L	15%	Ontario QMPLS	9.8 (3.9)	0.26
Fe	mcg/dL	20%	CLIA	4.2 (6.8)	0.34

PT Rules Studied

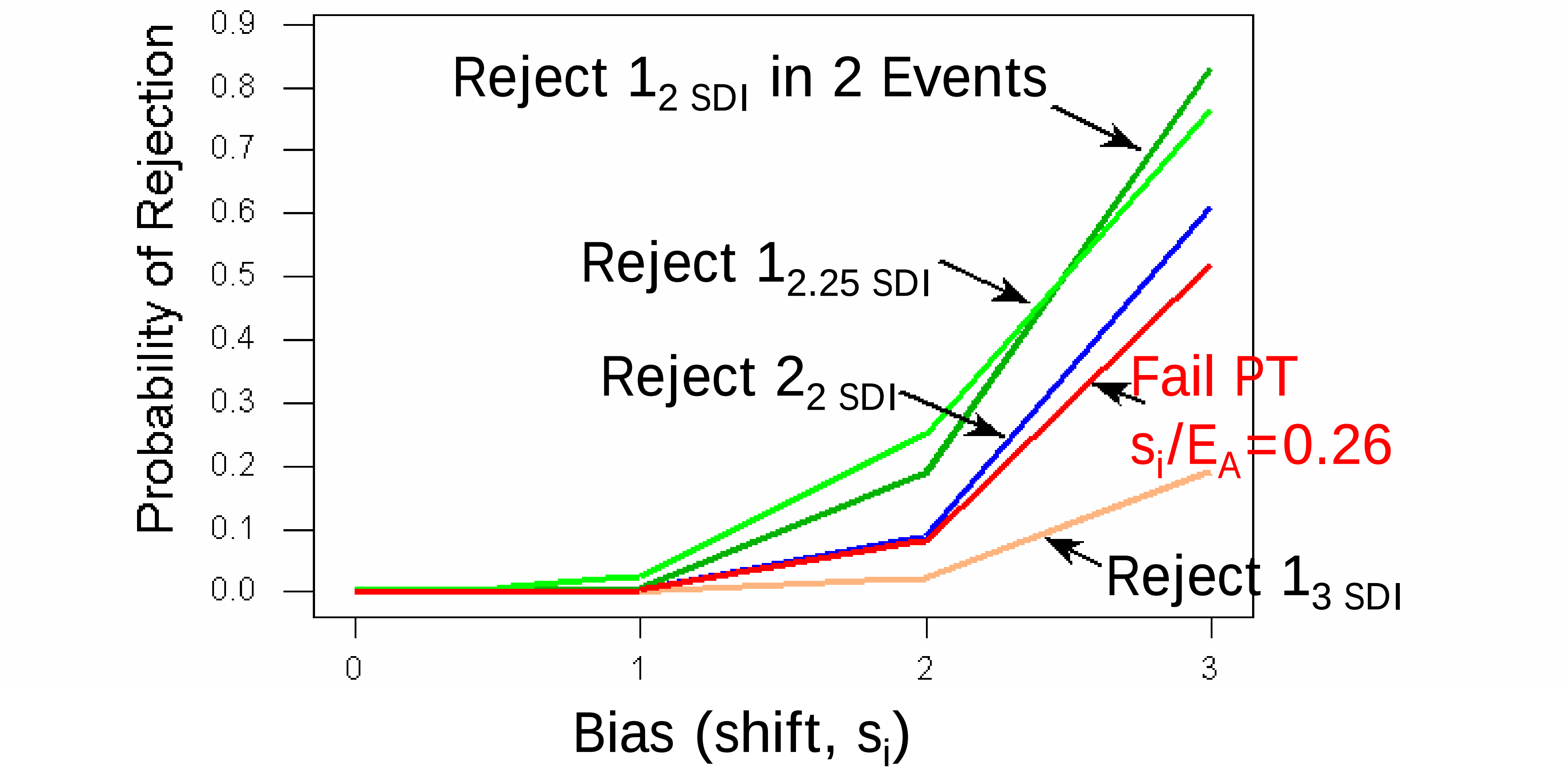
- Traditional: $1_{2\ SDI}$, $1_{2.25\ SDI}$,$1_{3\ SDI}$,$2_{2\ SDI}$,$1_{2\ SDI}$ in 2 PT events
- Range rules: $R_{3\ SDI}$, $R_{4\ SDI}$
- Mean rules: $\overline{X}_{1.0\ SDI}$ and $\overline{X}_{1.5\ SDI}$
- Percentage Rules: $1_{75\%\bullet EA}$ (one result or more has error >75% E_A) and $5_X\&1_{50\%\bullet EA}$ (all results on same side of mean and one or more has error >50% E_A)

Graphical examples of percentage rule failures:

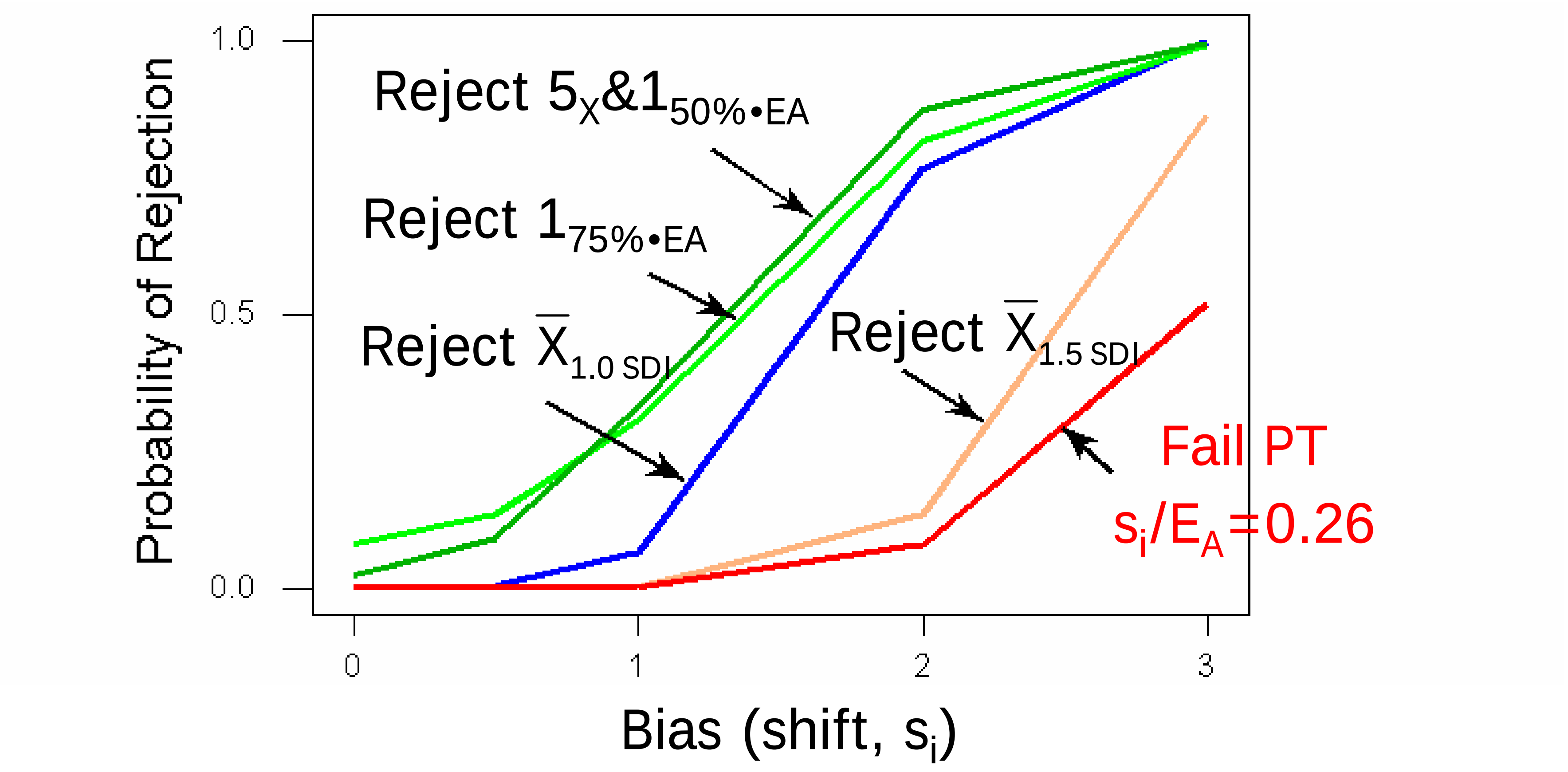
% Allowable Error

-100 -50 0 50 100
<+-----+-----+-----+-----+
: : 131 : : No failure (all $\leq$20%)
: : 1 : 1-1-1 **1** : $1_{75\%\bullet EA}$ Rule (One >75%)
: : : 1-3 **1** : $5_X\&1_{50\%\bullet EA}$ Rule (One >50%)

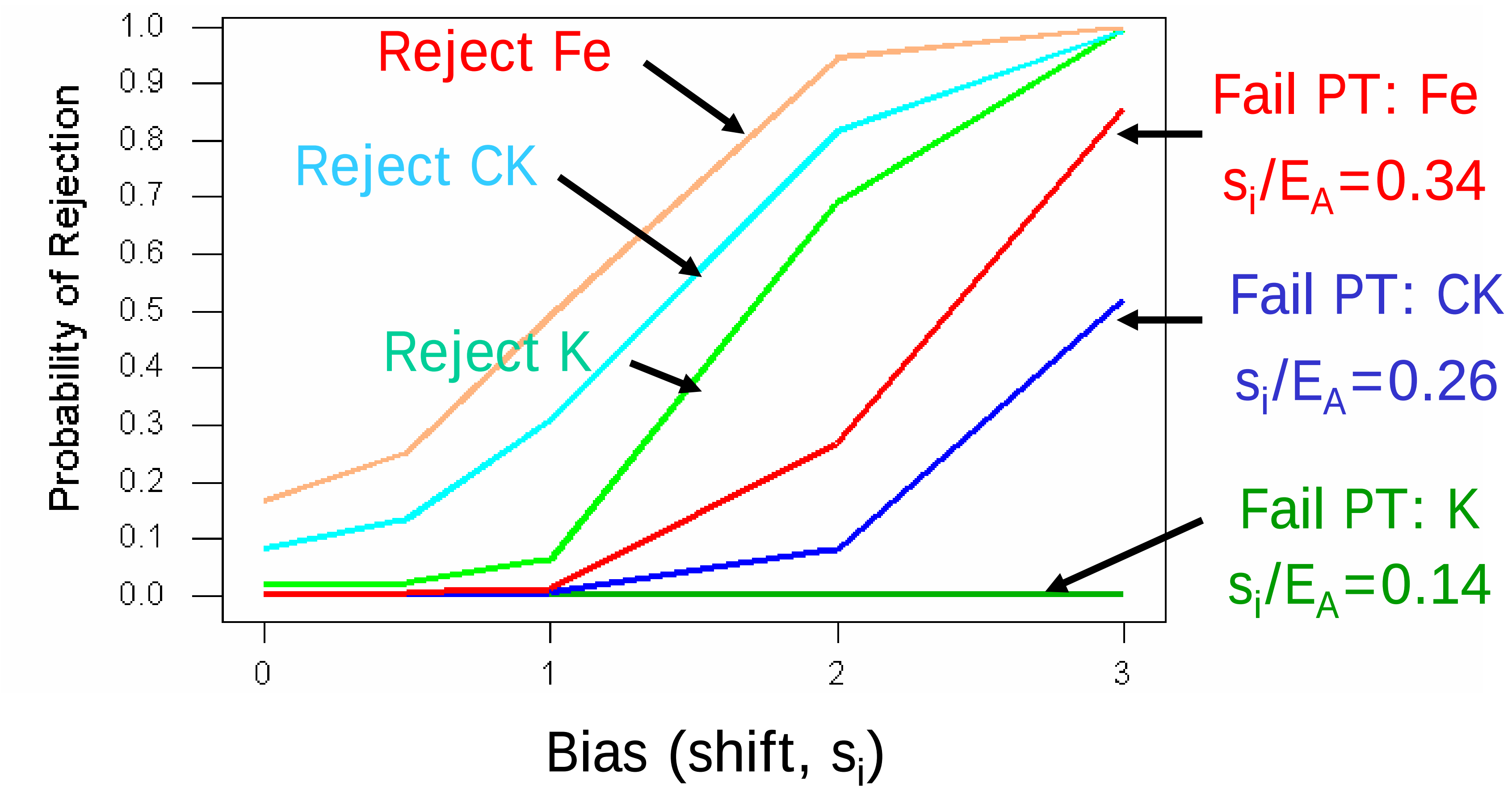
Results: Systematic Error, CK, Rejection by Traditional Counting PT Rules vs. PT Failure



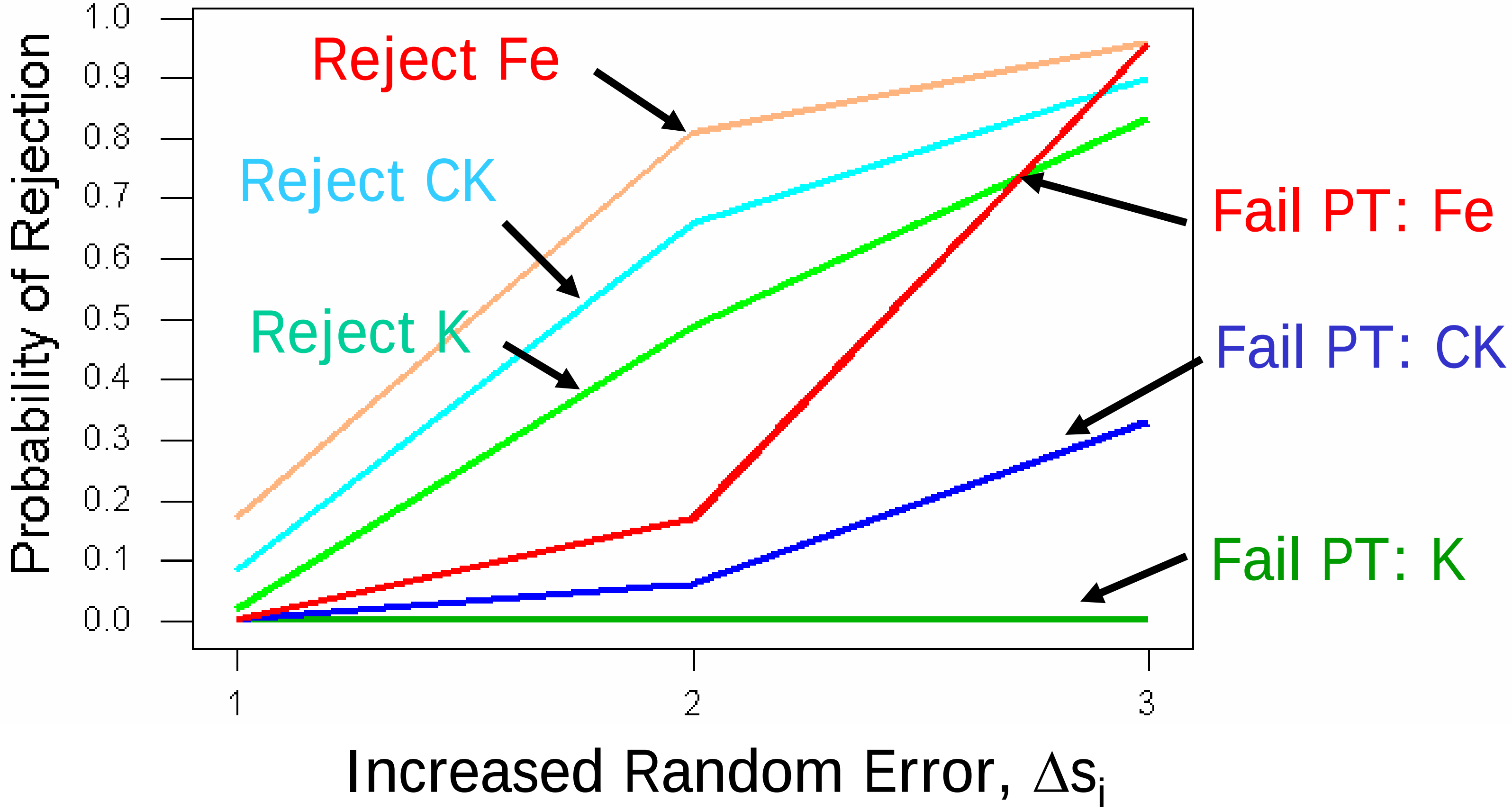
Systematic Error, CK, Rejection by Percentage Rules and Mean PT Rules vs. PT Failure



Systematic Error: $1_{75\%\bullet EA}/R_{4\,SDI}/X_{1.5\,SDI}$ Combination Rule Rejection vs. PT Failure



Random Error: $1_{75\%\bullet EA}/R_{4\,SDI}/X_{1.5\,SDI}$ Combination Rule Rejection vs. PT Failure



Observations

- **High quality methods (example, K)**
  - The probability of failing PT is nearly zero until errors are many times s_i.
  - Traditional counting rules perform well for interpreting PT data to detect significant errors.
- **Intermediate quality methods (ex., CK)**
  - Traditional counting rules lose power to detect errors.
- **Marginal quality methods (example, Iron)**
  - When significant errors occur, the probability of failing PT is higher than the probability that traditional counting rules will detect the errors.
  - Other rules with relatively high probability of false rejection must be used to detect errors.

Observations: continued

- **Sensitivity to errors** is decreased as the group standard deviation, s_g, increases relative to the internal standard deviation, s_i.

- **No single rule is effective** across the span of method quality and s_g/s_i values.

- **We recommend screening PT data** with the combination rule $1_{75\% \bullet EA}/R_{4\ SDI}/X_{1.5\ SDI}$

Investigating when only $1_{75\%\bullet EA}$ Rule Rejects

- Rejection may be caused by systematic error, random error, or false rejection

- **Test for systematic error** with sensitive follow-up rules:
  - 5_x &$1_{50\%\bullet EA}$ Rule is as sensitive as $1_{75\%\bullet EA}$ for systematic error
  - $\overline{X}_{1.0\ SDI}$ Rule is sensitive to systematic error
  - Both rules have low rates of false rejections.

- **Test for Random error**
  - $R_{3\ SDI}$ Rule is sensitive to random error
  - If no other rule failure, investigate further. The rate of false rejections is high for fair and marginal methods, but they are most prone to PT failure, and require vigilance.

Investigating $1_{75\%\bullet EA}/R_{4\ SDI}/X_{1.5\ SDI}$ Rule Rejections

- $\overline{X}_{1.5\ SDI}$ Rule rejection indicates systematic error. This can often be verified with peer QC data.
- $R_{4\ SDI}$ Rule rejection indicates random error.
- Both rules have low probability of false rejection.
- $1_{75\%\bullet EA}$ Rule responds to both systematic and random errors.

2001-2002 Experience with PT Combination Rule $1_{75\% \bullet EA}/R_{4\ SDI}/X_{1.5\ SDI}$

- 18 Rule rejects in 272 challenges, 7%
- 3 Were $1_{75\%\bullet EA}$ with $\bar{X}_{1.5\ SDI}$
- 2 Were $1_{75\%\bullet EA}$ with R_{4s}
- 1 Was R_{4s} only
- 10 Were $\bar{X}_{1.5\ SDI}$ only (4 with 5_x&$1_{50\%\bullet EA}$)
- 2 Were $1_{75\%\bullet EA}$ only - Probably real problems