The National Academies of
SCIENCES · ENGINEERING · MEDICINE

BOARD ON HUMAN-SYSTEMS INTEGRATION

DIVISION OF BEHAVIORAL AND SOCIAL SCIENCES AND EDUCATION

Basing Acceptance Criteria on Inferential Statistics versus Qualitative Analyses:

Lessons Learned from Medical Device User Interface Design

Ed Israelski PhD – AbbVie, Director HF

The Issues

- FDA expects Qualitative Evaluation of Safety Critical Use Scenarios for Medical Product User Interface Evaluations
  - Quantitative Acceptance Criteria (Usability Objectives) are not acceptable for product approval submissions
    - Example : 90% of first time minimally trained users will complete the task of programming an IV pump in < 5 minutes
    - Manufactures tried to game the system by claiming to beat the usability objectives, but not explain residual risk of observed task failures.
  - Totally different from FDA expectations for other quantitative acceptance criteria: sterility or drug stability that are based on inferential statistics

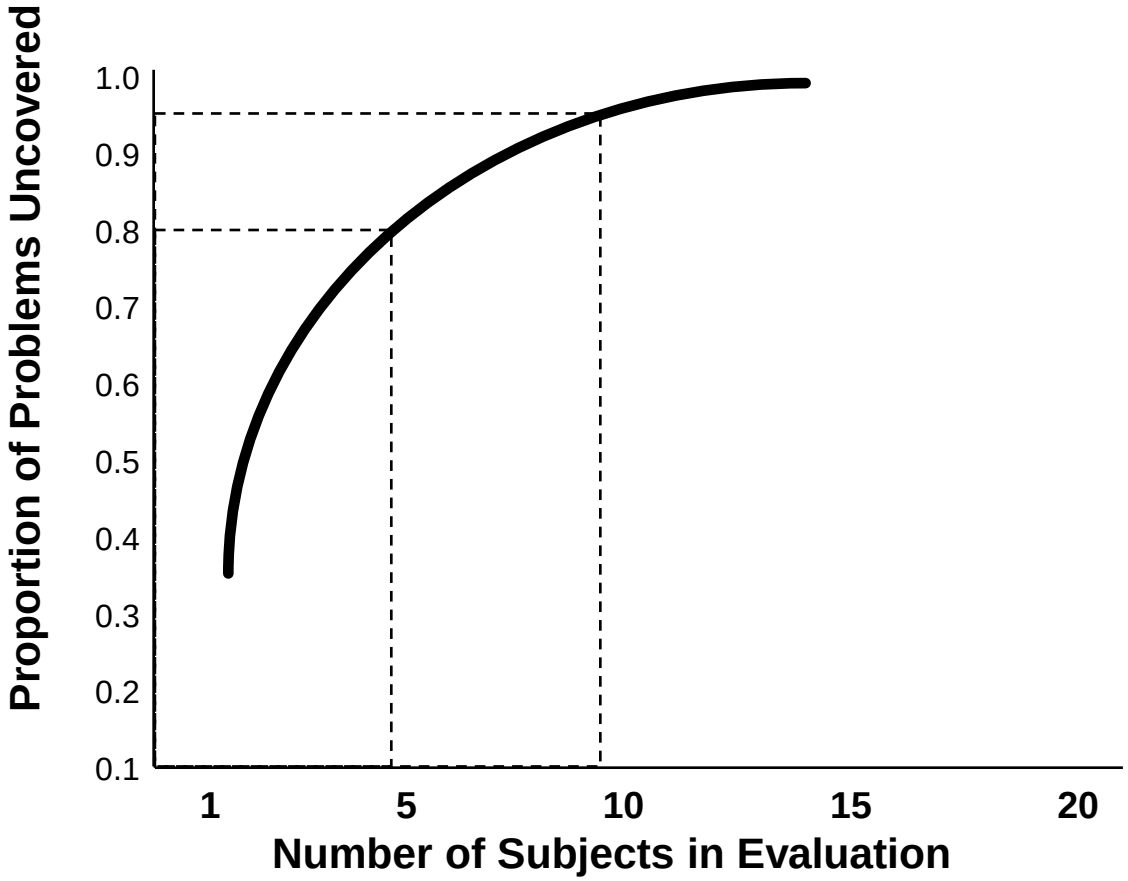
Issues

- Formative Usability Tests
  - Small sample sizes (5-8 per distinct user group)
  - Very qualitative (talk-aloud protocols)
- Summative (Validation) Usability Tests
  - N=15 per distinct group (FDA encourages >1 group)
  - Qualitative - may observe critical use errors with small samples, but FDA expects subjective probing
    - Question failures, close calls, use difficulties PLUS:
    - Probe for unobserved UI problems even with task success

Sample Sizes for Usability Testing

- Early Formative Testing
  - 5 to 8 is sufficient (per distinct user group)
  - High probability of catching >80% of usability defects
  - Includes exploratory, assessment and comparison tests
- Later Summative Validation Testing (Validates Usability)
  - 15 to 25 per major distinct user group (FDA requires 15 minimum)
  - **Safety Critical tasks** (Qualitative validation, explain/justify all failures)
    - Expect no failures (use errors) that have negative clinical impact
  - **Business Related Use Scenarios can be quantitative**
    - Statistical assessment of meeting Usability Objectives for business objectives
      - Via Classic Hypothesis testing (Inferential Statistics)
    - Set criteria and calculate 95% Confidence Interval and note if the upper confidence limit is above the target acceptance level
    - Or use ANSI/ASQC Z1.4-1993 sampling tables to set reject criteria values

Number of Subjects Needed in a Formative Usability Test



BOARD ON HUMAN-SYSTEMS INTEGRATION

Few Subjects Needed to Detect Usability Problems (Formative Usability Testing)

Underlying Use defect probability	Number of Test Participants													
	1	2	3	5	6	7	8	10	15	20	25	50	75	100
0.01	0.01	0.02	0.03	0.05	0.06	0.07	0.08	0.10	0.14	0.18	0.22	0.39	0.53	0.63
0.03	0.03	0.06	0.09	0.14	0.17	0.19	0.22	0.26	0.37	0.46	0.53	0.78	0.90	0.95
0.05	0.05	0.10	0.14	0.23	0.26	0.30	0.34	0.40	0.54	0.64	0.72	0.92	0.98	0.99
0.1	0.10	0.19	0.27	0.41	0.47	0.52	0.57	0.65	0.79	0.88	0.93	0.99	1.00	1.00
0.15	0.15	0.28	0.39	0.56	0.62	0.68	0.73	0.80	0.91	0.96	0.98	1.00	1.00	1.00
0.25	0.25	0.44	0.58	0.76	0.82	0.87	0.90	0.94	0.99	1.00	1.00	1.00	1.00	1.00
0.5	0.50	0.75	0.88	0.97	0.98	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
0.75	0.75	0.94	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
0.9	0.90	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

$R = 1 - (1 - P)^n$ Where **n** = number of test subjects

P = probability of a single test showing a problem or usability defect

R = Cumulative probability of detecting a usability problem

FDA Rationale for N=15 Sample Size

Table B-1. Percentage of Total Known Usability Problems Found in 100 Analysis Samples (Faulkner, 2003).

No. users	Min. % Found	Mean % Found	*SD*	*SE*
5	55	85.55	9.2957	.9295
10	82	94.69	3.2187	.3218
15	90	97.05	2.1207	.2121
20	95	98.4	1.6080	.1608
30	97	99.0	1.1343	.1051

Summary

- Qualitative Acceptance Criteria is less Burdensome for Manufacturers
- Powered studies as used in clinical trials would require 100's or 1,000's of test subjects
- Validation studies with 45-90 subjects plus extensive subjective probing has a decent chance of detecting critical use errors

Back Up Slides

Summative Usability Testing Sample Sizes

- Larger samples 15- 25 per distinct user group
  - Examples**: (Non Safety Critical Tasks Only)**
    - Objective is at least 90% complete a task the first time with no instructions
    - Hypothesis Testing:
      - Ho (Null) is Pass Rate >/= 90%
      - Ha (Alternative) is Pass Rate < 90%

Sample Size	Upper 95% Confidence Interval	Accept Ho (If # failed =/<)	Reject Ho
10	.912	3	4
15	.903	4	5
20	.929	4	5

Additional Sampling Plans – Summative Testing

Sampling Plan				
% Passing Objective	Sample Size	Accept	Reject	Upper 95% CL*
97%	10	1	2	99.5%
97%	15	2	3	97.6%
97%	20	2	3	98.2%
95%	10	2	3	96.3%
95%	15	2	3	97.6%
95%	20	3	4	95.8%
90%	10	3	4	91.3%
90%	15	4	5	90.3%
90%	20	4	5	92.9%
85%	10	4	5	85.0%
85%	15	5	6	85.8%
85%	20	6	7	86.0%
80%	10	4	5	85.0%
80%	15	6	7	80.9%
80%	20	7	8	82.3%
75%	10	5	6	77.8%
75%	15	7	8	75.6%
75%	20	8	9	78.3%
70%	10	5	6	77.8%
70%	15	8	9	70.0%
70%	20	10	11	69.8%

*Upper one-sided 95% confidence level for % Passing if the number of failures equals the Accept number

Need to Consider Type I & II Errors

Sampling Plan				Probability of Acceptance if True Population Passing Rate is***										
% Passing Objective	Sample Size	Accept	Reject	100%	99%	97%	95%	90%	75%	50%	25%	10%	Upper 95% CL*	Type I Error**
97%	10	1	2	1.00	1.00	**0.97**	0.91	0.74	0.24	0.01	0.00	0.00	99.5%	0.03
97%	15	2	3	1.00	1.00	**0.99**	0.96	0.82	0.24	0.00	0.00	0.00	97.6%	0.01
97%	20	2	3	1.00	1.00	**0.98**	0.92	0.68	0.09	0.00	0.00	0.00	98.2%	0.02
95%	10	2	3	1.00	1.00	1.00	**0.99**	0.93	0.53	0.05	0.00	0.00	96.3%	0.01
95%	15	2	3	1.00	1.00	0.99	**0.96**	0.82	0.24	0.00	0.00	0.00	97.6%	0.04
95%	20	3	4	1.00	1.00	1.00	**0.98**	0.87	0.23	0.00	0.00	0.00	95.8%	0.02
90%	10	3	4	1.00	1.00	1.00	1.00	**0.99**	0.78	0.17	0.00	0.00	91.3%	0.01
90%	15	4	5	1.00	1.00	1.00	1.00	**0.99**	0.69	0.06	0.00	0.00	90.3%	0.01
90%	20	4	5	1.00	1.00	1.00	1.00	**0.96**	0.41	0.01	0.00	0.00	92.9%	0.04
85%	10	4	5	1.00	1.00	1.00	1.00	1.00	0.92	0.38	0.02	0.00	85.0%	0.01
85%	15	5	6	1.00	1.00	1.00	1.00	1.00	0.85	0.15	0.00	0.00	85.8%	0.02
85%	20	6	7	1.00	1.00	1.00	1.00	1.00	0.79	0.06	0.00	0.00	86.0%	0.02
80%	10	4	5	1.00	1.00	1.00	1.00	1.00	0.92	0.38	0.02	0.00	85.0%	0.03
80%	15	6	7	1.00	1.00	1.00	1.00	1.00	0.94	0.30	0.00	0.00	80.9%	0.02
80%	20	7	8	1.00	1.00	1.00	1.00	1.00	0.90	0.13	0.00	0.00	82.3%	0.03
75%	10	5	6	1.00	1.00	1.00	1.00	1.00	**0.98**	0.62	0.08	0.00	77.8%	0.02
75%	15	7	8	1.00	1.00	1.00	1.00	1.00	**0.98**	0.50	0.02	0.00	75.6%	0.02
75%	20	8	9	1.00	1.00	1.00	1.00	1.00	**0.96**	0.25	0.00	0.00	78.3%	0.04
70%	10	5	6	1.00	1.00	1.00	1.00	1.00	0.98	0.62	0.08	0.00	77.8%	0.05
70%	15	8	9	1.00	1.00	1.00	1.00	1.00	1.00	0.70	0.06	0.00	70.0%	0.02
70%	20	10	11	1.00	1.00	1.00	1.00	1.00	1.00	0.59	0.01	0.00	69.8%	0.02

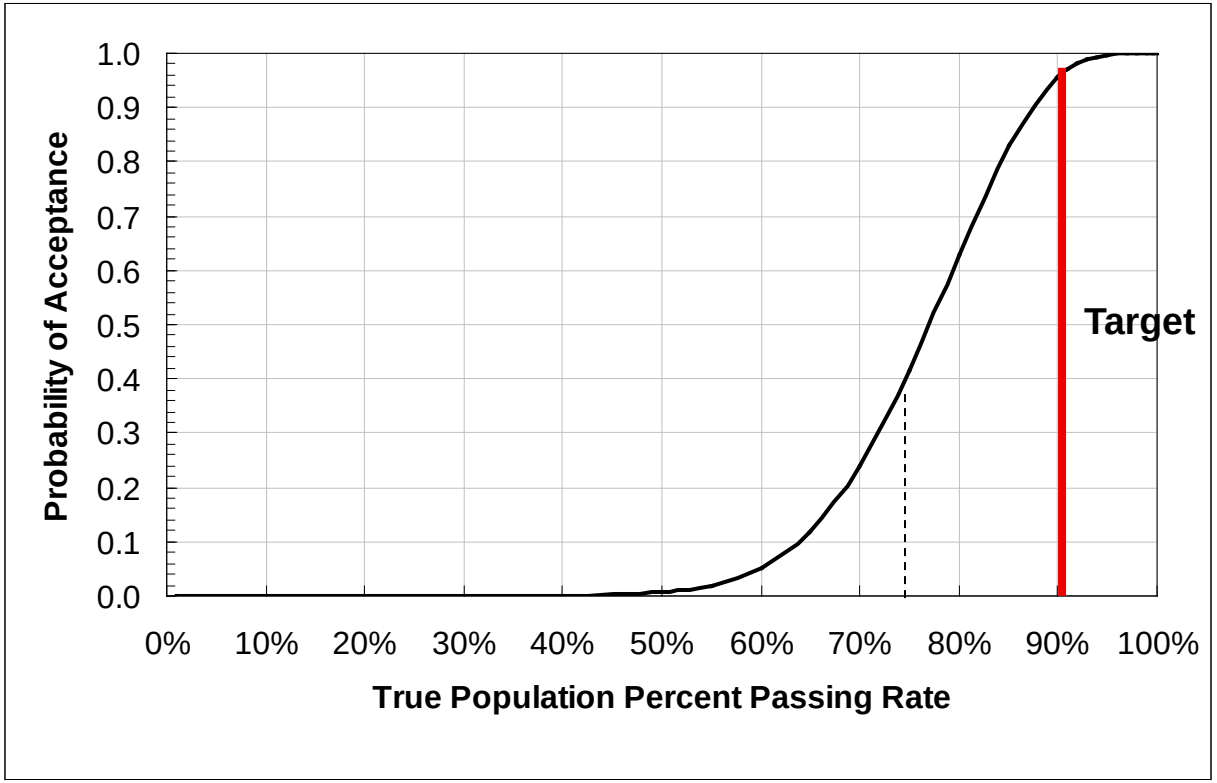
*Upper one-sided 95% confidence level for % Passing if the number of failures equals the Accept number

**Probability of falsely rejecting the null hypothesis of being better than the % Passing Objective

*** Type II error of falsely accepting the null hypothesis for true population values less than the % Passing Objective

Type II Error is a function of assumed passing rate for true population

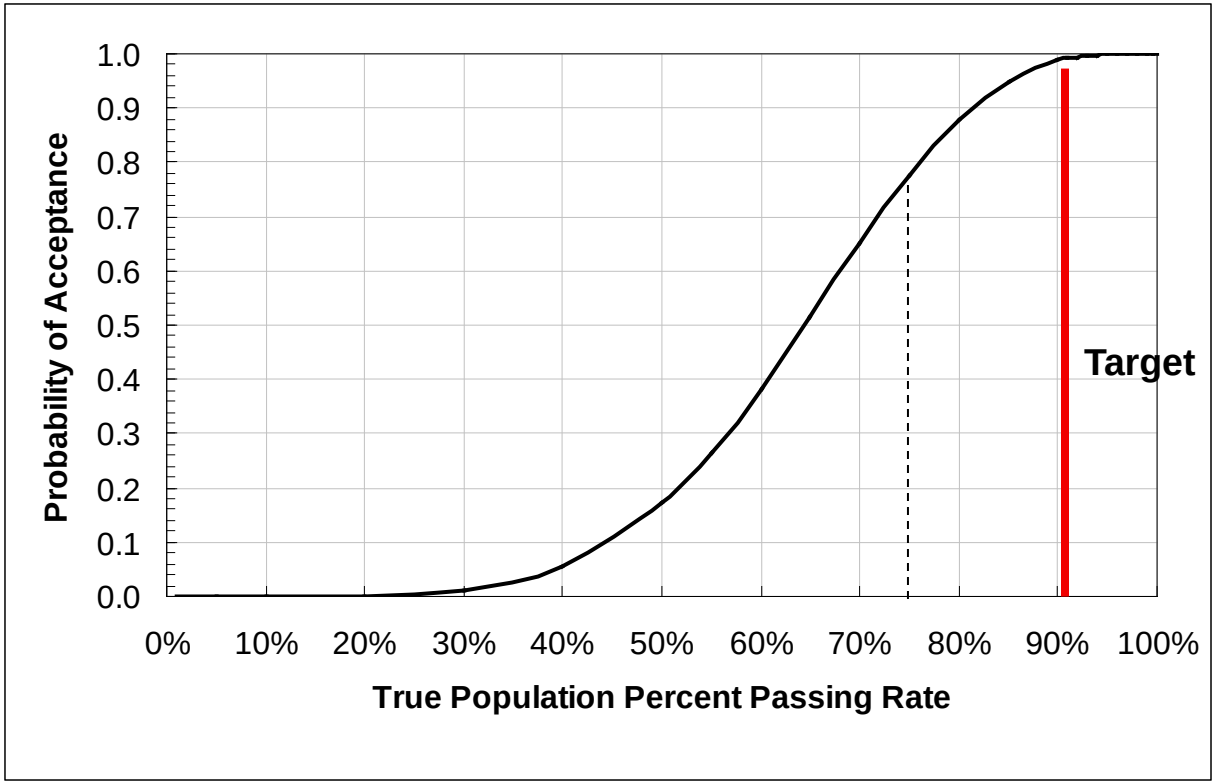
Type II Error
(Falsely Accepting
Null Hypothesis - Ho)



Ho: Passing Rate>90%
N=20 - reject Ho if 4 fail

Type II Error is a function of assumed passing rate for true population

Type II Error
(Falsely Accepting
Null Hypothesis - Ho)



Ho: Passing Rate>90%
N=10 - reject Ho if 3 fail

Conclusion: Smaller sample size has higher Type II error for any assumed True Population Passing Rate

Larger Sample Size for Power>80%

	Minimum Sample Size Necessary to Achieve 80% Power and (Accept Number) If True Population Passing Rate Is...												
% Passing Objective	100%	99%	97%	95%	90%	85%	80%	75%	70%	65%	60%	55%	50%
97%	*	*	*	*	66 (4)	36 (3)	21 (2)	11 (1)	9 (1)	8 (1)	7 (1)	6 (1)	5 (1)
95%	*	*	*	*	*	46 (5)	27 (3)	16 (2)	14 (2)	12 (2)	7 (1)	6 (1)	5 (1)
90%	*	*	*	*	*	*	78 (12)	40 (7)	25 (5)	18 (4)	13 (3)	11 (3)	8 (2)
85%	*	*	*	*	*	*	*	*	48 (11)	28 (7)	22 (6)	14 (4)	10 (3)
80%	*	*	*	*	*	*	*	*	*	56 (16)	35 (11)	21 (7)	17 (6)
75%	*	*	*	*	*	*	*	*	*	*	62 (21)	36 (13)	26 (10)
70%	*	*	*	*	*	*	*	*	*	*	*	67 (26)	39 (16)

* Requires a sample size > 100 or cannot be calculated

Summative Usability Testing Considerations

- **Optimistic Null Hypothesis e.g. Ho: Pass Rate>90% (N=15-25)**
  - Assumes all is well until data proves otherwise
    - Desirable outcome is to accept Ho
  - Assumes iterative design process that incrementally improves usability with each design iteration.
  - Cumulative sample size of successive formative usability tests can be large.
  - Can adjust ACCEPT number lower → Decreases Type II Error (Falsely accepting Ho) at the expense of Raising Type I Error (Falsely rejecting Ho)
- **Pessimistic Null Hypothesis e.g. Ho: Pass Rate<90% (N>50-80)**
  - Assumes all is NOT well until data proves otherwise
    - Desirable outcome is to reject Ho
  - Assumes no iterative design and use of formative usability tests
  - Can have test outcome of no task failures, but still cannot declare good usability