Overview

- Introduction
- Background
- The Purpose(s) of Evaluation
- Practical Measurement Guidelines
- Non-statistical Evaluation Designs
- Q & A
Who Are We?

- Trainers
- Instructional Designers/Developers
- Performance Consultants
- Human Resources Specialists/Mgrs
- Business Managers
- Researchers
- Who else?

Practitioners?

Background

- Less than 5% of ISPI article displays and chapters contain results data (Ogden Lindsley’s study)
- 1999 ISPI Think Tank Recommendations
- Practitioners vs. Researchers
- “Show Me The Money” I and II
- “Measurement Counts!” in Performance Express
- GOT RESULTS? Campaign
- Practical evaluation - “Single-case Design”
Why Do We Evaluate?

- To Validate
- To Hold or Be Held Accountable
- To Make Decisions

If you have data for making decisions, you’ve generally got the other two covered.

Brainstorm:
What Kinds of Decisions? What Kinds of Questions?

- Did it work? Did the intervention/program actually produce the desired results – behavior, accomplishments (job outputs), business results?
- How well? How BIG a difference did it make?
- How fast? At what rate/trend did the change take effect? Are there ANY trends or counter-trends?
- Better than another alternative? Does this type of intervention/program work better than some other type? Is it more cost-effective?
- Is it maintaining? Was our intervention a momentary blip, or did the results last?
What Can Practical Evaluation Achieve?

- Support for investment decisions
- Ability to predict performance outcomes
- Evaluation results that managers value and believe
- Meaningful evaluation as a routine of doing business
- New insights about what worked and why

Confidence in Results?

- Do we know our intervention CAUSED the effect?
- How BIG was the effect?
- How much BOUNCE or variability?

Would a scientist, engineer, accountant, or business manager BELIEVE our results?
What Do We Want to Avoid?

- “Results” that don’t predict performance outcomes
- Big expenditures of time and money
- Unacceptable interference with interventions
- The perception that “This is academic research.”
- Complex statistical designs (or complex statisticians!)
- Others?

Some Practical Measurement Guidelines

- Measure the right things.
  - Behavior
  - Accomplishments (job outputs)
  - Business Results
- Use standard countable units whenever possible.
- Include the time dimension.
- Recognize that rating scales are subjective – useful only when the desired accomplishments are things such as “People who think...” or “Customers who rate us as...” Then, count the people, don’t average their ratings.
The Foundation of Measurement is Counting with Standard Units!

We Can Count....

Behavior Influences
- Training
- Job Aids
- Incentives
- Feedback
- Ergonomics
- Tools
- Coaching
- Goal-setting
- Job Design
- Documents
- etc.

Behavior
- Explaining
- Asking
- Deciding
- Writing
- Speaking
- Finding information
- etc.

Job Outputs
- Proposals
- Problems solved
- Satisfied customers
- Buying decisions
- Signed contracts
- Repaired equipment
- etc.

Business Results
- Productivity
- Profits
- Market share
- Revenues
- Product sales
- Cycle time
- ROI
- Customers who say....
- etc.
Countable Units Corresponding to Gilbert’s Requirements

<table>
<thead>
<tr>
<th>Gilbert’s Requirements</th>
<th>Corresponding Countable Units</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Quality</strong></td>
<td></td>
</tr>
<tr>
<td>Accuracy</td>
<td>Count of accurate items</td>
</tr>
<tr>
<td></td>
<td>Count of inaccurate items</td>
</tr>
<tr>
<td>Class</td>
<td>Count of items in each category, rating, or class</td>
</tr>
<tr>
<td>Novelty</td>
<td>Count of items defined as novel, in separate categories if desired (e.g., new method, new product, new/better outcome, etc.)</td>
</tr>
<tr>
<td><strong>Quantity (or Productivity)</strong></td>
<td></td>
</tr>
<tr>
<td>Rate</td>
<td>Count of any behavior or accomplishment per unit of time (minute, hour, day, week, etc.)</td>
</tr>
<tr>
<td>Timeliness</td>
<td>Count of timely events or items</td>
</tr>
<tr>
<td></td>
<td>Count of untimely events or items</td>
</tr>
<tr>
<td>Volume</td>
<td>Count of items (as in “sales volume”), or Count of volume in units (e.g., liters, cubic yards)</td>
</tr>
<tr>
<td><strong>Cost</strong></td>
<td></td>
</tr>
<tr>
<td>Labor cost</td>
<td>Count of dollars spent on labor, by category</td>
</tr>
<tr>
<td>Material cost</td>
<td>Count of dollars spent on materials, by category</td>
</tr>
<tr>
<td>Management support cost</td>
<td>Count of dollars spent on management support, by category</td>
</tr>
</tbody>
</table>

100% Correct: As Good as It Gets???

?? “Overlearning” ??

Percent correct is not a measure of performance. It is a dimensionless quantity!
You Can’t Take Time Out of Performance!

The only upper limits are physical or environmental.

Count per minute is a true measure of performance.

Days

Count per Minute

Two Time Dimensions

Performance Occurs...

In Time

Over Time

Change can occur in Jumps (levels) and Turns (trends).

1. Count Per Min, Hr, Day, Wk, Mo, Qtr, Yr

2. Calendar Time (minutes, hours, days, weeks, months, years)
Levels can Jump while Trends Turn, for example.....

Flat before  Decelerating before  Accelerating before

Jump up  Turn up
Jump up  No Turn
Jump up  Turn down

And so on.....

So How Do We Tell What Happened?
Some Simple Evaluation Designs

- Just measure over time – “See if it’s changing.”
- Simultaneous comparison (groups, settings, etc.)
- Before -- After (baseline, change)
- Reversal (baseline, intervention, return to baseline)
- Repeated cases of Before -- After (replication)
- Before -- After at different times (“multiple baseline”)
  - Different individuals or groups
  - Different settings
  - Different measures
    - Behaviors
    - Accomplishments (job outputs)
    - Business Results

© 2002 Binder Riha Associates  4966 Wilshire Drive  Santa Rosa, CA 95404  (707)578-7850  10
**Just Measure Over Time**

- **Example:**
  - Measuring behavior from the beginning of a program until trainee meets criterion (i.e., measuring learning)

- **When Applicable:**
  - When you don’t have a baseline and you just want to see if things are going in the right direction, and if/when they achieve your target

- **Advantages:**
  - Better than nothing
  - If you measure the right thing, you'll be able to see if performance is moving in the right direction

- **Disadvantage:** Can’t be certain what caused the change.

---

**A Learning Chart**

Sample of individual practice card learning using 30-second measurement timings each day for two weeks. Correct responses (•) per minute accelerated by about x3.0 per week while errors plus skips (decelerated by about /2.6 per week over the course of the program.

By the end of the program, the trainee was making about 40 correct responses for every 1 error and was responding at over 80 correct per minute – a **fluent** level of performance on basic facts.
**Correct Responses Accelerate** (6 Trainees)

- Correct responses per minute for each member of the group.

**Simultaneous Comparison**

- **Example:** Measure productivity over time in multiple groups who receive different performance programs.
- **When Applicable:**
  - When you don’t have a baseline
  - If individuals or groups don’t have a history
  - When the controlling conditions are easy to identify, and you can be fairly certain that the groups/individuals are comparable.
- **Advantages:**
  - Better than nothing
  - When results are practically significant, this can be the easiest approach and does not require statistics to be convincing.
- **Disadvantage:** Always open to the criticism that the two individuals or groups were NOT comparable.
Group productivity measures in a customer call center with (•) and without (a) fluency-based new hire training.

The new-hire group trained using a fluency-based program accelerated call handling by about x1.4 per week for the two weeks after training, while those trained without the fluency program were flat (x1.0) or decelerated slightly (/1.1) over the two-week measurement period. By the end of the two week period, the fluency-trained representatives performed at about x1.6 (or 60%) higher than the nonfluency group.
**Why Use a Baseline?**

- **DEFINITION**: Repeated measures over time for a period BEFORE the intervention.

- **PURPOSE**: To provide a comparison with what happens after the intervention

- **Why REPEATED measures?** Makes us more confident that the effects weren’t just bounce or an extended trend.

**Before – After**

- **Examples**:  
  - New sales approach compared to previous approach  
  - New process compared to old process

- **When Applicable**:  
  - When you can obtain baseline measures  
  - When the results might not be reversible (e.g., learning)  
  - When multiple groups or baselines are not practical

- **Advantages**:  
  - Convincing if the results are large  
  - Can often use already-existing data

- **Disadvantage**: Open to the criticism that we don’t really know what caused the results.
Reversal to Baseline

- **Example**: Implementing an incentive system, then discontinuing it and continuing to measure.

- **When Applicable**:
  - Seldom applicable in business settings where positive results are desirable and managers would not want to reverse them.

- **Advantages**:
  - If the results return to baseline, it’s convincing that your intervention is what made the difference.

- **Disadvantages**:
  - Some effects will not reverse (e.g., learning)
  - Managers will not want to reverse positive outcomes
Repeated Cases of Before – After
(“Replication”)

- **Example:** Separately measuring sales regions that receive the same intervention at the same time.
- **When Applicable:** When you can gather data for separate individuals or groups, but can’t conduct pilot tests or stage sequential roll-outs.
- **Advantage:** Seeing the same effect many times is very convincing.
- **Disadvantage:** “Maybe it was the full moon!” – Open to the criticism that a calendar-related variable accounts for the difference (e.g., “March is always a better month…”)

Before - After at Different Times
(“Multiple Baseline”)

- **Examples:**
  - Pilot test with one or more groups before full implementation
  - Roll out programs to different groups spaced over months
  - Use implementation on a series of different “problems”
- **When Applicable:**
  - In any situation where you can implement sequentially across individuals, groups, locations, types of result, or improvement opportunities.
- **Advantages:**
  - Often a very practical fits with pilot-testing or staged rollout
  - When you get results, it is VERY convincing
- **Disadvantage:** Managers may insist on simultaneous implementations without any stages.
Switch to the Overhead Projector!

How Might YOU Apply These Ideas?
Some Parting Thoughts

- **Statistical significance is not enough.** We need interventions that are PRACTICALLY significant – *large enough to be obvious without statistics.*
- **Multiple baseline designs are practical:** They fit nicely with pilot testing, segmented implementation, and staged program roll-outs.
- **Stretch-to-fill graphs are confusing:** There are advantages to using standard charting methods.
- **We need more sharing:** The more we try these types of designs, and share the results with each other, the more we will learn and improve our practice.

References

**Publications**

**Links**
www.celeration.org – Web site of the Standard Celeration Society, a resource for use of Standard Celeration Charts and additional references about charting. Also, a charting list server (mostly academics and educators, with a few business people).
www-personal.umich.edu/~hinderer/scrdrefs.html – A rather unusual resource created by a physical therapist. Contains an extensive reference list on use of “single case designs,” the form of non-statistical evaluation designs upon which this presentation was based.
www.binder-riha.com – Binder Riha Associates’ site has “Helpful Resources” page.