



HIPPYUSA

Home Instruction for Parents of Preschool Youngsters

Measuring the Impact of HIPPY on Children: Pre-post Testing

Marsha M. Black, Ph.D.
University of South Florida
March 2004

Learning Goals

Readers of this module will be able to:

- **Describe pre-post testing.**
- **Discuss the limitations of the pre-post testing procedure for establishing cause and effect.**
- **Discuss the limitations of using developmental screening instruments for pre-post testing.**

Measuring the Impact of HIPPY on Children: Pre/Post Testing

HIPPY coordinators are increasingly under pressure to demonstrate that HIPPY is achieving its program goals and intended outcomes using some form of pre-post testing. Pre-post testing involves measuring growth in children's academic preparedness or progress, skills, knowledge, attitudes, or behaviors. Pre-post testing has also been used to measure change in child competencies in key skill areas of children's early development and learning. Funders and others believe that if a program is having an impact on its participants, the effect should be reflected as a positive change between participants' scores on a measure administered prior to participating in the program (the pretest) and their scores on the measure after completing the program or curriculum year (the posttest). Yet, this is not necessarily true. While pre-post testing can provide some information about changes in children's skills, it has serious limitations.

This learning module presents some basic information about pre-post testing including what the pre-post test procedure can show about program effectiveness and the limitations of this approach as a research design. Two case studies will be presented illustrating the kinds of information that can be learned from using the pre-post testing procedure using developmental screening instruments. It is hoped this information will help clarify some of the issues surrounding pre-post testing, and that you will be in a better position to make decisions that are in the best interests of your HIPPY program.

Pre-Post Testing

Difference or gain score = posttest score minus pretest score

What is pre-post testing? Pre-post testing requires collecting and analyzing data across two observations: baseline (beginning of program) and at a later point, at the end of the program year or when the program has been operational long enough for possible change to occur. The posttest score minus the pretest score is called a “difference” score (also called a “gain score” or “change score”). Individualized child tests and teacher/parent reports/ratings are common data collection methods used in pre-post testing. The pre-post testing procedure involves selecting the appropriate outcome, selecting a measuring instrument, deciding when to administer the measure, and calculating the difference or gain score.

Limitations of Pre-Post Testing as a Study Design for Establishing Cause-Effect Relationships

Though one group of participants may increase their scores on a particular measure from pretest to posttest, this does not necessarily mean the increased scores can be attributed only to participation in the HIPPIY program. If the evaluation design had included a representative comparison group (children who were similar to the HIPPIY children but were not in the HIPPIY program) who also took the pre-post tests, then the posttest change of the HIPPIY children’s scores beyond that which occurred in the control group could be attributed to participating in HIPPIY.

Extraneous Variables

An extraneous variable is any variable other than the treatment variable that, if not controlled, can affect the study outcomes.

In order to establish with a high level of confidence a cause and effect relationship, extraneous variables (variables other than participation in the HIPPIY program) must be controlled or minimized. Consider the following examples:

- Interventions that extend over long periods of time provide the opportunity for other events to occur in the lives of participants besides receiving the intervention. For example, as a result of parents

enrolling in GED classes, the family environment becomes more educationally focused and supportive of learning.

- If the interval between the pretest and posttest is very long, an external event (parent may have enrolled the child in a preschool program or in a tutoring program) or maturation (child's natural growth) could be affecting the amount of gain between the two tests.
- If the same test is used as both a pretest and posttest, there could be a testing effect. That is, the children might show an improvement simply because of their experience with the pretest.
- If the pretest and posttest are very different, the learning gain may be due to the change in the nature of the measuring instrument.
- Participants who take a pretest may perform better in the program than those who do not take the pretest because the pretest increases awareness, stimulates learning, and/or enhances preparation for program activities.

Increasing the Usefulness of the Pre-Post Test Design

The degree of usefulness of the pretest-posttest design increases to the extent that some of the extraneous factors discussed in the preceding paragraph can be controlled or minimized. There are several ways this can be accomplished:

- If the pretest and posttest measure is a standardized, norm or age-referenced tests the standardization sample can be considered an acceptable substitute for a control group (the demographics of the standardization sample are described in test administration manuals and norm tables are provided for comparison data).
- Using the same instrument as a pretest and posttest or using two properly equated tests such as comparable forms of the same test safeguards against instrumentation concerns (tests with comparable forms are listed in the module Commonly Used Assessment Instruments).
- Selecting a sample of HIPPY children that are representative of all HIPPY children participating in the program addresses the some of

the concerns regarding selection bias.

Despite the limitations associated with using the one group pretest-posttest evaluation design (i.e., there is no comparison group in the study) to establish cause-effect relationships, pre-post testing can be useful for some purposes. Gain scores (the posttest score is greater than the pretest score) do reflect a positive change or growth from pretest to posttest. This suggests the program may be having some positive impacts on its participants. It is important to remember there is no perfect research design. Unanticipated events can and do arise in research and evaluation studies that can affect outcomes. However, well-designed evaluation activities facilitate the collection of quality data that can be useful and informative and usefulness is one of the goals of conducting program evaluation activities.

Developmental Screening Instruments

- **Purpose:**
 - Screen to identify children in need for special services or supports

- **Advantages:**
 - Can be administered quickly
 - Items easy to understand

Many preschool and kindergarten programs have mandated formal developmental screenings for all children they serve in order to identify children with developmental delays at an early age. Screening instruments are designed to identify children who should be referred for further assessment to determine the need for special services or supports.

Developmental Domains

Most developmental screening instruments have items that are clustered within five domains:

Personal-Social Domain - abilities and characteristics that facilitate children engaging in positive and meaningful social interactions. The behaviors

measured include adult interaction, expression of feelings/affect, self-concept, peer interaction, coping, and social role.

Adaptive Domain - self-help skills and task-related skills. Self-help skills are those behaviors that enable the child to become increasingly more independent in daily living skills such as feeding, dressing, and personal toileting needs. Task-related skills involve the child's ability to pay attention to specific stimuli for increasingly longer periods of time, to assume personal responsibility for his or her actions, and to initiate purposeful activity and follow through appropriately to completion. Behaviors measured include attention, eating, dressing, personal responsibility, and toileting.

Motor Domain - gross motor development (large muscle movement and control) and fine motor development (hand and finger skills; and hand-eye coordination). Behaviors measured include muscle control, body coordination, locomotion, fine muscle, and perceptual motor.

Communication Domain - Understanding and using language to communicate for various purposes. Behaviors measured include a child's reception and expression of information, thoughts, and ideas through verbal and nonverbal means.

Cognitive Domain - skills and abilities that are conceptual in nature. Abilities measured include perceptual discrimination, memory, and reasoning. Tasks include comparison among objects based on physical features (color, shape, size) and properties (weight); sequencing events; putting together parts of a whole; grouping and sorting similar objects and identifying similarities and differences among objects based on common characteristic.

Using A Developmental Screening Instrument as a Pre-Post Test

Because they are already being administered to children and because they are quick and easy to administer, it is tempting to use developmental screening instruments as pre-post measures for program evaluation. However, developmental screening instruments have limitations when used to measure child progress.

Characteristics of Developmental Screening Instruments

- Few items, not very sensitive
- Discontinuous item distribution
(Items are clustered in the middle with few items on either ends of the continuum)
- Limited usefulness for measuring child progress (pre-post measure)
- Limited usefulness for comparing groups of children

The most commonly used developmental screening instruments are standardized, normed instruments meaning there are instructions for administering items, and information is available on how typical children score on the test. Because they are designed for administration to large numbers of children as the first stage in a program of assessment, they contain a limited number of items and can be administered quickly. These few items do not measure the entire range of achievement, and thus these instruments are of limited usefulness in measuring child progress over time (a pre-post measure).

For some development screening instruments, total raw scores in each domain are compared to pre-established cutoff points. Scores above the cutoff point mean the child is progressing as expected for his/her chronological age and scores below the cutoff point may indicate a need for follow-up. Other developmental screening instruments provide instructions for converting raw scores to age-equivalent scores. Converting raw scores to age-equivalent scores for two groups of children provides a comparison of the difference in average age-equivalent scores for the two groups at each point in time. This provides a more meaningful interpretation of raw scores

than simply the percentage of children above and below the pre-established cutoff points.

The limitations of developmental screening instruments can best be illustrated through the experiences of two programs that used a screening instrument as a pre-post test to document the effectiveness of the HIPPY program. (See examples on the following pages.)

Program I - Ages and Stages Questionnaire

Characteristics of Ages and Stages Questionnaire

- Standardized and normed instrument
- Parent-report instrument
- Few items; general in nature

Domains

- Communication
- Gross motor
- Fine motor
- Problem Solving

The Ages and Stages Questionnaire is a parent-report instrument with 19 questionnaires with 30 developmental items that can be administered at regular intervals from 4 months to 60 months (5 years) of age. Reading level ranges from 4th to 6th grade and items measure behaviors across five developmental domains. As with any typical developmental screening instrument, there are few items in each domain and the items are general in nature. For example, suppose we have a 36 month old boy, and we want to measure his performance in the Personal-Social Domain. As shown below, there are only six items in this domain to measure a complex construct that includes such skills as:

- Ability to play with other children.
- Self-help skills such as feeding and dressing oneself.
- Self-awareness and personal knowledge such as knowing name and age.

Questions in Personal-Social Domain For child 36 months old

- Does your child use a spoon to feed himself with little spilling?
- Does your child push a little shopping cart, stroller, or wagon, steering it around objects and backing out of corners if he cannot turn?
- When she is looking in a mirror and you ask, “Who is in the mirror?” does your child say either “Me” or her own name?
- Can your child put on a coat, jacket, or shirt by himself?
- Using these exact words, ask your child, “Are you a girl or a boy?” does your child answer correctly?
- Does your child take turns by waiting while another child or adult takes a turn?

Scoring: Raw Scores and Cutoff Points

Yes = 10 Sometimes = 5 Not yet = 0

- Total raw score
- Cutoff points: pass/fail
- Example: Cutoff point for Personal-Social domain for child 36 mos. is 38.7.

0	5	10	15	20	25	30	35	40	45	50	55	60
4	4	4	4	4	4	4	4	4	4	4	4	4

Scoring the Ages and Stages Questionnaire involves determining if the child performed the specific behavior described in each of the items within a specific domain. Ten points are scored if “yes”, 5 points if “sometimes” and zero points if “not yet”. Scores are then summed across all of the items and compared to a cutoff point. **A cutoff point is a number that identifies the point for pass/fail for each domain for each chronological age.** Scores falling above the cutoff point mean the child is progressing as expected for his/her developmental age. Scores below the cutoff point mean a child may need further diagnostic assessment. The only information you can determine from a comparison of raw scores to cutoff points is whether or not the child is above or below the predetermined cutoff points for Time 1 and Time 2.

Because the test provides numerical scores, we are tempted to perform calculations using scores. We may want to average all the scores together and say that, on average, children scored “X” points. But there is no way to interpret what “X” means. We may want to compare the performance of two children. For example, say the first child scored 30 and the second child scored 60. Because 60 is twice 30, we may want to say that the second child performed twice as well as the first child who only scored 30. **But we cannot say how much better one child does than another. All we can say is that one child scored above the cutoff and “appears to be doing well in this area at this time” and the second child “may need additional assessment services.”** In addition, we also cannot group scores. The only information we can extract from using this instrument as a pre-post test is the number and percent of children above and below the cut off score at each point in time.

Battelle Developmental Inventory

Characteristics of the Battelle Developmental Screening Instrument

- A standardized, norm-referenced test
- An individually-administered test but also uses observation and parent report data
- Five developmental domains
 - Adaptive
 - Motor
 - Communication
 - Cognitive
 - Personal-Social
- Also available as a full assessment battery
- Requires trained tester

The second program used the Battelle Developmental Screening Inventory. Like the Ages and Stages Questionnaire, the Battelle is a standardized, norm-referenced test. The Battelle has five developmental domains, and though there are a few more items in the domains, the range covered by the items is still limited. Unlike the Ages and Stages, the Battelle is also available as a full assessment battery and requires a trained person to administer the test because basal and ceiling rules apply to each

domain. The basal rule is used to determine the level of item difficulty below which the child would get all of the items correct. The ceiling rule is used to determine the level of item difficulty above which the child would get 0 on all of the items. The Battelle utilizes observation and interview data to evaluate a child's skills or to gather additional information about a child's performance in a specific area.

Scoring: Raw scores, Cutoff points, Age Equivalent scores

Similar to the Ages and Stages Questionnaire, the criteria for scoring a child's performance are presented with each item, using the following three-point system:

- 2 points = The child responds according to the specified criterion
- 1 point = The child attempts an item but cannot meet the specified criterion
- 0 points = The child cannot or will not attempt an item, or the response is an extremely poor approximation of the desired behavior

- Raw scores compared to cutoff point: pass/fail
- Age-equivalent scores can be compared with child's chronological age

Age equivalent scores indicate the age at which a raw score is average.

Advantages:

- Places the obtained score in a developmental context
- Provides information that is easily understood
- Can compare the relative performance of two groups of children taking the same test
- Meets the requirements of some local, state, and federal programs that require age scores for eligibility and/or funding purposes.

Example: if the average score of a 10 year old child on a particular domain is 15 out of 25 then any child obtaining a score of 15 items out of 25 would have an age equivalent score of 10.

Determining a child's age equivalent score in a specific domain:

- Consult the Age Norms (Age-Equivalent Score) in Months for the domain you are interested in. (Note: There are separate age equivalent scores for each domain.)
- Calculate the child's total score in the domain of interest.
- Locate the range of values that contains the Total Score.
- The AE will be listed opposite the range of values that contains the Total Score.

Scoring Summary Sheet

- Scores across all items in one domain are summed and compared to a cutoff point indicating pass/fail

Domain	Raw Score	Cutoff Score	Pass/Fail	Age Equivalent Score
Personal-Social	37	29	Pass	60 months
Adaptive	28	27	Pass	63 months
Communication	19	20	Fail	33 months
Motor	30	26	Pass	49 months
Cognitive	25	24	Pass	44 months
Total Score	139	135	Pass	46 months

In the example above, you will notice there is a column for recording the age equivalent score for each domain. This is a major advantage of the Battelle Developmental Screening Instrument compared to the Ages and Stages Questionnaire.

Unlike the program that used Ages and Stages, this second program compared the performance of two groups of children enrolled in an early intervention preschool program. There were 15 children in the HIPPIY group and 15 children in a comparison group. The Battelle was used as a pre-post measure and was administered as a pretest in September and as a posttest in February.

Determining the *average age equivalent* scores for a group of children

- Sum all of the Total Scores in each domain for children in one group and divide by the number of children in that group.
- Repeat process for the second group.
- Subtract the posttest value from the pretest value for each group of children.
- Compare the difference.
- Children whose performance improved from pretest to posttest will have higher posttest values.

The results from the comparison of age equivalent scores for the HIPPY and non-HIPPY groups of children were as follows:

	Pre	Post	Difference
HIPPY	48.0	61.2	13.2
Non-HIPPY	47.2	55.2	8.0

As noted above, the difference in average composite scores between HIPPY and non-HIPPY children is +5.2 indicating the performance of HIPPY children exceeded that of the non-HIPPY children. However, even though there was a comparison group of non-HIPPY children, the pre-post test instrument was a developmental screening instrument, which is not suitable for measuring child progress over time (see the discussion in the Developmental Screening Instruments section of this learning module).

Having a comparison group enables you to measure how much change has occurred, but not how much of the observed change can be attributed to participating in the HIPPY program. Why is this? Because developmental screening instruments have limited item difficulty (most children would pass most of the items) and item sampling (items typically don't measure the entire range of a child's abilities).

In conclusion, there is a difference in the conclusions that were drawn from the two analysis:

Ages and Stages Questionnaire:

Can say:

“A significant proportion of HIPPY children scored above the cutoff point on the Ages and Stages at Time 1 and at Time 2.” Or “Of the ____% of children who scored below the cutoff point at Time 1, ____% scored above the cutoff point at Time 2.”

You have a number and a percentage for each of the five domains, but you cannot make statements about “how much” improvement was made.

Battelle Developmental Screening Instrument:

Can say:

“For the first and second BDI test, the average age-equivalent scores for the HIPPY and non-HIPPY groups were lower than their chronological ages suggesting poor performance relative to others in their age group. There was improvement in age equivalent scores for both groups of children. However, for children in the HIPPY group, there was an average difference of 13.2 points between the first and second BDI, while for the non-HIPPY group, the average difference was 8 points.”

You can test this difference with a statistical procedure that will tell you whether the observed difference is significant beyond chance.

Conclusion

It is important to remember that there is no “perfect” instrument for use as pre-post test. All instruments will have advantages, disadvantages and limitations on what the results will tell you. For further information on specific instruments that may be used for pre-post testing, consult the module entitled “Commonly Used Assessment Instruments”. The more knowledgeable you are about these limitations, the better positioned you will be to make decisions that are in the best interest of your program.

