

5 Adequacy of evaluation design



Ideally, evaluations of interventions should be systematic, comprehensive and use rigorous scientific controls, such as randomised trials and sufficient statistical power, to find meaningful program effects (Sanders 2003). Some existing reviews of program evaluations have developed standards, grades or levels of evidence for early childhood interventions, based on certain criteria. These categories are used as a means of reporting the rigour of the evaluation design (for example, Mrazek and Brown 2002).

Evidence rating system

The evidence rating system adopted in this report aims to provide information on a number of fundamental research design elements. The elements included in this review are:

- *Appropriate evaluation design methodology.* Evaluations (including cost-benefits analyses) require an appropriate control or comparison group. This can be achieved either by randomly assigning participants to be in the intervention or control group, or by selecting a group of participants that are matched to the intervention group on a number of characteristics such as gender and age (matched comparison group).
- *Pre-intervention data.* For matching intervention and control groups, and to detect change as a result of implementation, it is necessary to collect baseline information.
- *Intermediate follow-up and long-term follow-up.* To determine whether the intervention has had any short-term and/or long-term effects, outcome data should be regularly collected on the intervention and comparison groups. Ideally, follow-up should continue for a number of years.
- *Representative sample of participants in the evaluation.* To ensure that an evaluation is representative of the intervention it is evaluating, the evaluation sample must be representative of the whole sample that received the intervention.
- *Low attrition at follow-up and non-random attrition.* Attrition in regard to evaluation integrity refers to the number of participants that could not be included in the immediate or long-term follow-up. Attrition is generally deemed to be acceptable if it is no more than 10 per cent per follow-up time point. Therefore, in a sample of 100, no more than ten participants could be lost at each follow-up time point.
- *Adequate statistical power.* To ensure that an evaluation is statistically adequate, the case-to-variable ratio used in an analysis needs to be considered. A minimum of five participants for every one characteristic measured is standard.
- *Reliable measures.* The integrity of an evaluation is enhanced if the tools used to measure outcomes are standardised (that is, have known psychometric properties) and widely used.
- *Appropriate choice of measures.* In making decisions about how outcomes are to be measured, serious consideration must be given to the measures used. A measure that does not adequately assess what evaluators want it to assess will compromise the integrity of the evaluation.
- *Appropriate analytic approach.* This criterion refers to the use of appropriate statistical techniques. This is necessary to ensure that the findings are reliable.

The presence or absence of each design element is recorded in Tables 1-5 below. Full details of the intervention evaluations and outcomes are provided in Appendix 2.

Adequacy of cluster 1 evaluations

All evaluations in cluster 1 included a representative sample of participants. Most used reliable measures, made appropriate choices about measures and used appropriate analytic approaches. Four of the six interventions (Perry, CPC, High/Scope and PIDI) included an appropriate control or comparison group and four (Perry, Head Start, High/Scope, PIDI) collected pre-intervention data. Half of the interventions had follow-up data (Perry, CPC, High/Scope).

The evaluation integrity of three interventions in cluster 1 was very good, with all three interventions containing nine of the ten research design elements (Perry, CPC, High/Scope). The evaluation integrity of one intervention (Saginaw) was very poor, containing only two of the research design elements; while the evaluation integrity of the remaining two interventions (Head Start, PIDI) was moderate (six design elements). These details are illustrated in Table 1.

Adequacy of cluster 2 evaluations

All but one of the evaluations in cluster 2 (SHELLS) contained an appropriate control or comparison group. All of the evaluations included pre-intervention measures. SHELLS and Baby HUGS did not collect follow-up data, while the remaining evaluations included at least intermediate follow-up data. Half of the evaluations did not have adequate statistical power and half did not use reliable measures.

The evaluation integrity of one intervention (Elmira PEIP) was excellent, reflecting all ten of the design elements. One intervention (SHELLS) had very poor evaluation integrity (one design element present) while the evaluation integrity of the remaining six interventions was moderate to good. These details are illustrated in Table 2.

	Perry	Head Start ¹	CPC	High/Scope	Saginaw ²	PIDI
Includes appropriately-matched comparison group or randomised control design methodology	√	x	√ ³	√	x	√ ⁴
Pre-intervention (baseline) data available	√	√	x	√	x	√ ⁵
Intermediate follow-up (i.e. collected up to two years after the intervention period)	√	x	√	√	x	x
Long-term follow-up (i.e. collected more than 2 years after the intervention period)	√	x	√	√	x	x
Representative sample of participants included in the evaluation ⁶	√	√	√	√	√	√
Low attrition at longitudinal follow-up (not more than 10 per cent per data point) and attrition not systematic	√	NA	√	√	NA	NA
Adequate statistical power for analyses	x ⁷	√	√	x ⁸	√	√
Reliable measures	√	√	√	√	x	NR
Appropriate choice of outcome measures	√	√	√	√	x	√
Appropriate analytic approach	√	√	√	√	x	√
Number of evaluation design elements present	9/10	6/10	9/10	9/10	2/10	6/10
√=design element present x=design element not present NA= not applicable (for example, no longitudinal follow-up) NR=not reported (insufficient information published to determine whether design element present/absent)						
1 Numerous evaluations of Head Start have been conducted. Given limited time frames, this review focuses on a large-scale national evaluation, however it must be noted that this is not necessarily representative of all evaluations of Head Start. 2 Evaluations of this program examine whether or not the intervention group has achieved the objectives set out by the program. 3 Although participants in the Chicago CPC were self-selected, the intervention and control groups did not differ on a number of characteristics at the beginning of the intervention. 4 Participants in the program were self-selected. 5 On most measures. 6 However, those receiving the program were often not representative of the general population (i.e., mostly African American children). 7 Numerous analyses were conducted on a small sample, meaning that some findings may be significant due to chance. 8 Numerous analyses were conducted on a small sample.						

Adequacy of cluster 3 evaluations

All of the evaluations of interventions in cluster 3 included appropriate control or comparison groups, a representative sample, adequate statistical power, reliable measures and chose appropriate outcome measures.

	PEIP	PCDC	HIPPY ⁹	Healthy Start	EEP	SHELLS	Baby HUGS	Project 12-ways
Includes appropriately-matched comparison group or randomised control design methodology	√	√	√	√	√	x	√	√ ¹⁰
Pre-intervention (baseline) data available	√	√	√	√	√	√	√	√
Intermediate follow-up (i.e. collected up to two years after the intervention period)	√	√	√	√	x	x	x	√
Long-term follow-up (i.e. collected more than 2 years after the intervention period)	√	√	x	x	√	x	x	x
Representative sample of participants included in the evaluation	√	NR	√	√	√	x	x	√
Low attrition at longitudinal follow-up (not more than 10 per cent per data point) and attrition not systematic	√	x	√	x	√	NA	x	NA ¹¹
Adequate statistical power for analyses	√	x	x	√	√	x	x	√
Reliable measures	√ ¹²	√	x	x	√	x ¹³	√	x
Appropriate choice of outcome measures	√	√	√	x	√	x ¹⁴	√	x
Appropriate analytic approach	√	√	√	√	NR	x	√	√
Number of evaluation design elements present	10/10	7/10	7/10	6/10	8/10	1/10	5/10	6/10

√=design element present
x=design element not present
NA= not applicable (for example, no longitudinal follow-up)
NR=not reported (insufficient information published to determine whether design element present/absent)

9 Not all of the evaluations reviewed were adequate.
10 Participation in the program was not random.
11 The evaluation was conducted by examining Department of Children and Family Services files only.
12 Although some of the measures used are questionable as they rely solely on maternal report.
13 Outcomes were assessed largely by parent report only.
14 Child outcomes were not assessed (but are planned for future evaluations).

	New Hope	FTP	TPDP
Includes appropriately-matched comparison group or randomised control design methodology	√	√	√
Pre-intervention (baseline) data available	√	√	x
Intermediate follow-up (i.e. collected up to two years after the intervention period)	√	√	x
Long-term follow-up (i.e. collected more than 2 years after the intervention period)	x	√	√
Representative sample of participants included in the evaluation	√	√	√
Low attrition at longitudinal follow-up (not more than 10 per cent per data point) and attrition not systematic	√	x	√
Adequate statistical power for analyses	√	√	√
Reliable measures	√	√	√
Appropriate choice of outcome measures	√	√	√
Appropriate analytic approach	√	√	x ¹⁵
Number of evaluation design elements present	9/10	9/10	7/10

√=design element present
x=design element not present
NA= not applicable (for example, no longitudinal follow-up)
NR=not reported (insufficient information published to determine whether design element present/absent)

15 The significance level used was 0.10.

Table 3 shows that the evaluation integrity of two of the interventions was very good, with both evaluations containing nine of the ten design elements (New Hope, FTP). The evaluation integrity of the remaining intervention (TPDP) was good, containing seven design elements.

Adequacy of cluster 4 evaluations

Most of the evaluations in cluster 4 included a representative sample and chose appropriate outcome measures, while two-thirds of the evaluations included an appropriate control or comparison group and two-thirds used reliable measures. For most of the other design elements, approximately half contained each design element. Attrition in the evaluations was acceptable in only four of the evaluations (Abecedarian, IHDP, Incredible Years, ECEAP) and were not applicable in half of the interventions due to the lack of longitudinal follow-up.

The evaluation integrity of three interventions was very good, with all evaluations containing nine of the ten design elements (Abecedarian, IHDP, Incredible Years). Two interventions (Sure Start and NEWPIN) had very poor evaluation integrity, with each intervention containing only one design element. However, more comprehensive evaluations of Sure Start are pending. The evaluation

	Early Head Start	Abecedarian	IHDP	Syracuse	SESS	Even Start	CCDP	Incredible Years	ECEAP	BBBF	Sure Start ¹⁶	NEWPIN
Includes appropriately-matched comparison group or randomised control design methodology	√	√	√	√	√	√	x	√	x ¹⁷	√	x	x
Pre-intervention (baseline) data available	x ¹⁸	√	x	x	√	√	√	√	x	√	x	x
Intermediate follow-up (i.e. collected up to two years after the intervention period)	x ¹⁹	√	√	x	x	√	√	√	√	x ²⁰	x	x
Long-term follow-up (i.e. collected more than 2 years after the intervention period)	x	√	√	√	x	x	√	x	√	x ²¹	x	x
Representative sample of participants included in the evaluation	√	√	√	√	√	√	√	√	√	√	x	x
Low attrition at longitudinal follow-up (not more than 10 per cent per data point) and attrition not systematic	NA	√	√	x	NA	x	x	√	√	NR	NA	NA
Adequate statistical power for analyses	√	x	√	x	√	x	√	√	√	√	x	x ²²
Reliable measures	√	√	√	√	√	√	√	√	x	NR ²³	x	NR
Appropriate choice of outcome measures	√	√	√	√	x	√	√	√	√	√	√ ²⁴	√
Appropriate analytic approach	√	√	√	NR	√	√	x	√	x ²⁵	√	x	NR
Number of evaluation design elements present	6/10	9/10	9/10	5/10	6/10	7/10	7/10	9/10	6/10	6/10	1/10	1/10
√=design element present x=design element not present NA= not applicable (for example, no longitudinal follow-up) NR=not reported (insufficient information published to determine whether design element present/absent)												
16 A comprehensive evaluation is pending. 17 The intervention and comparison groups differed quite dramatically on level of poverty. 18 Minimal baseline data was collected. 19 An intermediate follow-up is planned. 20 Longitudinal analyses are planned. 21 Longitudinal analyses are planned. 22 One evaluation did, however this evaluation focused primarily on service use, rather than outcomes. 23 The measures were not described adequately enough to make a judgement. 24 The planned outcome measures are appropriate. 25 Much of the evaluation focused on comparing groups within the intervention group, rather than comparing the intervention and comparison groups.												

	Triple P	PAT	Cuyahoga
Includes appropriately-matched comparison group or randomised control design methodology	√	x	x
Pre-intervention (baseline) data available	√	x	√
Intermediate follow-up (i.e. collected up to two years after the intervention period)	√	√	√
Long-term follow-up (i.e. collected more than 2 years after the intervention period)	x	x	x
Representative sample of participants included in the evaluation	√	√	√
Low attrition at longitudinal follow-up (not more than 10 per cent per data point) and attrition not systematic	x	x	√
Adequate statistical power for analyses	√	x ²⁶	√
Reliable measures	x ²⁷	√	x
Appropriate choice of outcome measures	√	√	x
Appropriate analytic approach	√	x	x
Number of evaluation design elements present	7/10	4/10	5/10
√=design element present x=design element not present NA= not applicable (for example, no longitudinal follow-up) NR=not reported (insufficient information published to determine whether design element present/absent) 26 Not for the all of the analytic procedures used. 27 The measures used were largely parent report. Some observations were also conducted.			

integrity of the remaining seven evaluations was moderate to good (five to seven design elements). These details are illustrated in Table 4.

Adequacy of cluster 5 evaluations

All three of the evaluations in cluster 5 contained an intermediate follow-up and a representative sample, however none of them contained a long-term follow-up. In addition, attrition was high in all but one evaluation (Cuyahoga) and only Triple P included an appropriate control group and used an appropriate analytic approach.

As shown in Table 5, the evaluation integrity of Triple P was good (seven design elements); the evaluation integrity of PAT was poor (four design elements); and the evaluation integrity of Cuyahoga was moderate (five design elements).

Relative adequacy of evaluations across clusters

It is difficult to make any firm distinctions between clusters, given the great variability in evaluation integrity within clusters. With the exception of cluster 5, each cluster contained evaluations with very good integrity, while all clusters except cluster 3 contained evaluations with very poor to poor integrity.

One design element that warrants further discussion is the use of reliable measures. Regardless of cluster, most of the evaluations included some objective measures, as well as parental reports. Although parent reported measures have their merit, and are usually the most expedient way of data collection, they are subjective by nature. Objective measures are therefore needed to corroborate parental reports.