Outline model Impact Evaluation toolkit

Author: Robert Davies

October 2020

NIDA Network for Information and Digital Access

Contents

1. Impact and its evaluation: an introduction

There are various definitions of Impact. In its framework of international development, the OECD Development Assistance Committee (DAC), 2010 defines impact as, "Positive and negative primary and secondary long-term effects produced by an intervention, whether directly or indirectly, intended or unintended.'[1]

The terms Impact Evaluation (IE) and Impact Assessment (IA) are sometimes used interchangeably, and it may not be possible to avoid such terminological overlaps. In this summary, IE is used to describe a specific type and purpose (a 'discipline') within the general field of evaluation, and an IA to describe specific cases or instances with a defined methodology.

IE has strong roots in the fields of medicine and health involving the principles of comparison, under controlled conditions and statistical analysis of the possibility of error. Drawing on the work of Archibald Cochrane in the 1960s and 70s, Randomised Controlled Trials (RCT) involving control groups, randomization, and blinding are seen as the core methodology in assessing clinical therapies. However, RCTs are far from straightforward. Reflecting a more general trend toward evidence-based policies, research and practice, interest in IE has spread to many fields in education, science, social science and humanities. In these contexts, an 'intervention' might therefore be a small project, a large programme, a collection of activities, or a policy.

Impacts are usually understood to occur later than, and as a result of, intermediate *outcomes*. For example, achieving the intermediate outcomes, say, of improved access to land and increased levels of participation in community decision-making might occur before, and contribute to, the intended final impact of improved health and well-being for women. However, the distinction between outcomes and impacts can be relative, depending on the stated objectives of an intervention. Some impacts may be 'emergent' as things proceed, and thus, cannot be predicted.

Evaluation, by definition, answers questions about quality and value. This is what makes evaluation more useful and relevant than simply measuring indicators or summarising observations and stories. IE can have formative, summative, advocacy-oriented and other purposes. In any impact assessment, it is important to define first what is meant by 'success' (e.g. quality, value). This can lead to defining specific evaluation criteria with different levels of performance (or standards), deciding what evidence will be gathered and how it will be synthesized to reach defensible conclusions. Those defined by OECD-DAC include: Relevance, Effectiveness, Efficiency, Impact and Sustainability.

Evaluation relies on a combination of facts and values to judge the merit of an intervention. Like any other evaluation, an impact evaluation should be planned formally and managed as a discrete project, with decision-making processes and management arrangements clearly described from the beginning of the process. Participatory approaches are often favoured, raising the question: who should be involved and why? Purpose, objectives and target audience help determine the need for rigour, methodology (and cost).

IA frequently assess changes in KAB: knowledge, attitudes, behaviours. A 'theory of change' may explain at the outset how activities are understood to produce a series of results that will contribute to achieving the ultimate intended impacts. This can continue to be revised over the course of evaluation, should either the intervention itself or the understanding of how it works change. The IA might confirm the theory of change or it may suggest refinements based on the analysis of evidence. Failure to achieve the final intended impacts might be due to theory failure rather than implementation failure. A Theory of Change is, however, helpful in guiding causal attribution in an Impact Assessment.

Determining causal attribution, i.e. a*scription of a causal link between observed (or expected to be observed) changes and a specific intervention,* is often seen as requirement for calling an activity Impact Evaluation and the challenges of this are a major issue in IE. It is not, however, necessary that changes are produced solely or wholly by the intervention under investigation, and it is understood that other causes or contextual factors may also be taken into account ('external' or 'extraneous' factors'), for example prior knowledge, motivation, and interest.

[1] http://www.oecd.org/sti/inno/What-is-impact-assessment-OECDImpact.pdf

2. Ontological framework

An ontological framework for IA in Science Literacy would form a key element of a toolkit for those planning implementations. The basic structure for such a framework would require an implementer to consider and define, at least, the following (some definitions of technical terms are provided at the end of this document).

Impact Assessment research aspect	Some options for consideration
Science subject area	
Geographic area	
Educational delivery model	Formal, non-formal, programme-embedded
Target sectors	
Delivery institution	
Research design	Experimental, quasi-experimental, non-experimental
Audience and comparison samples	RCT, control group, comparison group
Testing procedure	Pre- and post-test, longitudinal test
Sampling procedure	Random, cluster, convenience, multi-stage, systematic
Sample size	
Data collection approach	Quantitative, qualitative, mixed-method
Data collection methods	Written, online or phone survey, focus/discussion group, interviews, observation, case studies
Data collection tools	Likert scales, multiple choice or dichotomous questions, tests
Data analysis tools	Excel, SAS, SPSS, Stata
Statistical tests for data variance, significance and reliability	t-tests, t-based confidence intervals, ANOVA, ANCOVA, MANOVA, chi-squared, Wilcoxon signed rank test, Cronbach's Alpha reliability test

Fig: outline ontology for impact assessment planning

Definitions and examples of many of these aspects and options are provided in sections 4 and 5 below.

3. Impact Evaluation methods

There are many different methods for collecting data and analysing in IE. A 'mixed method' evaluation is the systematic integration of quantitative and qualitative methods. This helps to overcome the weaknesses inherent in each method when used alone. The credibility of evaluation findings can be increased when information from different data sources converge to deepen the understanding of the effects and context of an intervention, for example:

- Descriptive questions ask about how things are and what has happened, the initial situation and how it has changed, the activities of the intervention, participant characteristics, and the implementation environment.
- Causal questions ask whether or not, and to what extent, observed changes are due to the intervention itself rather than to other factors, including other programmes and/or policies.
- Evaluative questions ask about the overall conclusion as to whether an intervention can be considered a success, an improvement or the best option.

Using a combination of these strategies can usually help to increase the strength of the conclusions.

Three design options that address causal attribution are referred to as:

- Experimental designs – which construct a control group through genuine random assignment, along the lines proposed for RCT
- Quasi-experimental designs – which construct a comparison group by other means.
- Non-experimental designs – which look systematically at whether the evidence is consistent with what would be expected if the intervention were producing the impacts or whether other factors could provide an alternative explanation.

These different approaches are important when deciding what will be considered credible by the intended user of the evaluation or by partners or funders. The design options all need significant investment in preparation and early data collection and cannot be done if an impact assessment is limited to a short exercise conducted towards the end of intervention implementation. **IE is best addressed as part of an integrated monitoring, evaluation and research approach that generates a range of evidence to inform decisions.**

Other key aspects in designing valid IA with acceptable results involve sampling methods and sizes, the timing and longitudinality of pre- and post-testing, data collection and analysis and statistical methods for analysing variance and significance. The importance of interpretation in creating narratives from results should not be forgotten and can be assisted by well-defined original goals and objectives for an IA. The extent and sophistication of impact assessment in the science literacy field, its academic documentation and evident momentum over two decades exceeded the expectations held by the study team at the outset of the *impact of Science Literacy delivery methods - what works?* research. Impact assessment methodologies in Science Literacy emulate those apparent in other fields and focus on research design, validity, analysis and interpretation, covering key aspects, such as:

- Experimental, Quasi-Experimental, Non-Experimental design;
- quantitative, qualitative or 'mixed methods' approaches;
- sampling and data analysis, visitor responses;
- statistical significance; and
- causality and understanding extraneous variables.

A temporary tendency to prefer qualitative data in some institutional contexts over 'hard' quantitative data has in more recent years been supplanted by a preference for 'mixed-method' approaches. Sometimes research findings are characterised by weaknesses such as descriptive designs, small samples over narrow time frames, reliance on student or instructor perceptions, researcher developed instruments, and multiple interventions. This can lead to calls for wider focus, larger samples, greater longitudinality and sometimes efforts to implement RCTs. However, resource constraints remain a significant factor in this direction.

In summary, proposals for methodological improvement to impact assessment frequently relate to:

- increasing the longitudinality of measurement;
- multiple time-points to measure the persistence of effects;
- quasi-experimental designs, control groups and bigger samples to increase validity;
- careful empirical study design and objective analysis;
- more direct methods of measurements;
- focusing on explorative learning processes;
- better quantification of learning, beyond subjective perceptions of teachers and taught;
- benefits of using pre- and post-visit data on the same visitors for direct comparison;
- using different methods in combination for 'triangulation';
- combining quantitative and qualitative methods;
- pre-defined data scales and reporting rules;
- proper weighting through appropriate power calculations;
- using appropriate materials (e.g. pictography in health literacy) with specific audiences to determine literacy on particular issues; and,
- comparing different types of visual narratives which may rely on different cognitive mechanisms.

4. Definitions of Impact Assessment terms

4.1. Research types and designs

Qualitative research is a type of social science research that collects and works with non-numerical data and that seeks to interpret meaning from these data that help understand social life through the study of targeted populations or places, typically focused on the micro-level of social interaction. Methods of qualitative research include: observation and immersion, interviews, open-ended surveys, focus groups, content analysis of materials, and oral history.

Quantitative research is the process of collecting and analysing numerical data and employing statistical operations to determine causal and correlative relationships between variables. It can be used to identify large-scale trends, find patterns and averages, make predictions, test causal relationships, and generalize results to wider populations.

Mixed-methods research uses a combination of quantitative and qualitative data to evaluate something, during development (formative or iterative evaluation), how well it worked (summative evaluation), or to identify needs for improvement. Combining quantitative and qualitative approaches can balance out the limitations of each method, provide stronger evidence and more confidence in findings and give more granular results than each individual method. Drawbacks include: greater complexity, a requirement for greater expertise to collect and analyse data and interpret results, rather than using a single method.

Experimental research designs are based on a clear hypothesis; the purpose of the research is to confirm or refute the validity of the hypothesis. They have an independent variable, a dependent variable, and a control group. A **quasi-experimental design** is an empirical study, almost like an experimental design, but without random assignment. Quasi-experimental designs typically allow the researcher to control the assignment to the treatment condition but using some criterion other than random assignment (e.g., an eligibility cut-off mark). In some cases, the researcher may have control over assignment to treatment condition. In **Non-experimental** designs the researcher cannot control, manipulate or alter the predictor variable or subjects, but instead, relies on interpretation, observation or interactions to come to a conclusion. Typically, this means relying on correlations, surveys or case studies, and cannot demonstrate a true cause-and-effect relationship. Non-experimental research tends to have a high level of external validity, meaning it can be generalized to a larger population.

A **randomized controlled trial** (RCT) is a type of scientific (often medical) experiment that aims to reduce certain sources of bias when testing the effectiveness interventions by randomly allocating subjects to two or more groups, treating them differently, and then comparing them with respect to a measured response. One group (the experimental or treatment group) receives the intervention being assessed, while the other (usually called the control group) does not. The trial may be **blinded,** meaning that information which may influence the participants is withheld until after the experiment is complete with the intention to reduce or eliminate some potential sources of experimental bias.

Pre-test post-test designs are one of the simplest methods of testing the effectiveness of an intervention.

Two-Group Control Group Design is a useful way of ensuring that an experiment has a strong level of internal validity**.** The principle involves randomly assigning subjects between two groups, a test group and a control. Both groups are pre-tested, and both are post-tested, the ultimate difference being that only one group experienced the intervention.

A **longitudinal study** is a research design that involves repeated observations of the same variables (e.g., people) over short or long periods of time (i.e., uses longitudinal data), which can be either an observational study, or structured as longitudinal randomized experiments. They may be retrospective (looking back in time) thus using existing data, or prospective (requiring the collection of new data). Longitudinal studies do not require large numbers of participants.

4.2. Sampling methods

Random sampling is a technique in which each sample has an equal probability of being chosen and is meant to be an unbiased representation of the total population. If for some reason, the sample does not represent the population, the variation is called a sampling error.

Cluster sampling means a method of surveying a population based on groups naturally occurring in a population. The population within a cluster should ideally be as heterogeneous as possible, but there should be homogeneity between clusters. Each cluster should be a small-scale representation of the total population. The clusters should be mutually exclusive and collectively exhaustive. A common motivation for cluster sampling is to reduce the total number of interviews and costs given the desired accuracy.

Multistage sampling is a sampling method that divides the population into groups (or clusters) for conducting research. Significant clusters of the selected people are split into sub-groups at various stages to make it simpler for primary data collection.

Convenience sampling consists of research subjects who were chosen for a study because they could be recruited easily. One disadvantage is that the subjects may not be representative of the population the researcher is interested in studying. An advantage of convenience sampling is that data can be collected quickly and for a low cost. Convenience samples are often used in pilot studies, through which researchers can refine a research study before testing a larger and more representative sample.

Stratified sampling is a method in which the total population is divided into smaller groups or strata, based on some common characteristics in the population data, such as age, gender, nationality, job profile, educational level etc. After dividing the population into strata, the researcher randomly selects the sample proportionally. Stratified sampling is used when the researcher wants to understand the existing relationship between two groups.

Systematic sampling means that the researcher first randomly picks the first item or subject from the population, then selects each n-th subject from the list. The results are representative of the population unless certain characteristics of the population are repeated for every n-th individual, which is highly unlikely. The procedure involved in systematic random sampling is very easy and can be done manually. The process of obtaining the systematic sample is much like an arithmetic progression.

4.3. Data collection tools

Surveys are a data gathering method that is utilized to collect, analyse and interpret the views of a group of people from a target population. Survey methodology is guided by principles of statistics from the moment of creating a sample, or a group of people to represent a population, up to the time of the survey results' analysis and interpretation. **Questionnaires** are utilized in various survey methods and may be administered in person or by phone or can be self-administered (this is now commonly done online, as in the form of web surveys).

Likert scales are used where respondents are asked to indicate the level of agreement or disagreement with certain statements. They provide a series of answers to choose from to one or more questions, ranging from one extreme attitude to another, normally with a moderate or neutral option. They are popular because they are one of the most reliable ways to measure opinions, perceptions, and behaviours.

Multiple choice questions involve choosing the correct answer from a list of possible answers

Dichotomous questions can have two possible answers, usually used in a survey to ask for Yes/No, True/False or Agree/Disagree answers, for clear distinction of qualities, experiences or respondent’s opinions.

Focus groups use group dynamics to obtain shared experiences of people with similar characteristics; differing from **interviews**, where the focus is on individuals. They can be used before, during, or after an intervention when there is a need to gain a breadth of understanding of the thoughts and experiences of people, or there are open-ended questions to address.

Observation (or field research) is a social research technique that involves the direct observation of phenomena in their natural setting. There are different types of observational methods such as: Controlled Observations, Naturalistic Observations, and Participant Observations, with different degrees of influence on participant behaviour.

4.4. Statistical software

SPSS (Statistical Product and Service Solutions)**, SAS and Stata** are examples of software available for retrieving, managing, editing data from a variety of sources and performing statistical analysis on it, applicable in social sciences as well as other scientific fields

4.5. Statistical methods

Confidence intervals. A confidence interval is a way of using a sample to estimate an unknown population value. For estimating the mean, two types of confidence intervals that can be used: z-intervals and t-intervals**. t-tests** (see below) are used when the Population standard deviation is unknown and the original population normal or sample size greater than or equal to 30.

Regression. Regression is a statistical method that attempts to determine the strength and character of the relationship between one dependent variable (usually denoted by Y) and a series of other variables (known as independent variables). **Linear regression** is a basic and commonly used type of predictive analysis. The overall idea of regression is to examine two things: (1) does a set of predictor variables do a good job in predicting an outcome (dependent) variable?; and (2) Which variables in particular are significant predictors of the outcome variable?

ANOVA. Analysis of variance (ANOVA) is an analysis tool used in statistics that splits an observed aggregate variability found inside a data set into two parts: systematic factors and random factors. The systematic factors have a statistical influence on the given data set, while the random factors do not. Analysts use the ANOVA test to determine the influence that independent variables have on the dependent variable in a regression study.

ANCOVA. Analysis of covariance (ANCOVA) is a general linear model which blends ANOVA and regression. ANCOVA evaluates whether the means of a dependent variable (DV) are equal across levels of a categorical independent variable (IV) often called a treatment, while statistically controlling for the effects of other continuous variables that are not of primary interest, known as covariates (CV) or nuisance variables.

MANOVA. multivariate analysis of variance (MANOVA) is a procedure for comparing multivariate sample means. As a multivariate procedure, it is used when there are two or more dependent variables and is often followed by significance tests involving individual dependent variables separately. MANOVA is a generalized form of univariate ANOVA although distinguished by using the covariance between outcome variables in testing the statistical significance of the mean differences.

chi-squared independence test. This is a procedure for testing if two categorical variables are related in some population e.g. if education level and marital status are related for all people in some country.

t-test. A t-test is a type of inferential statistic used to determine if there is a significant difference between the means of two groups, which may be related in certain features. It is mostly used when the data sets, like the data set recorded as the outcome from flipping a coin 100 times, would follow a normal distribution and may have unknown variances. A t-test is used as a hypothesis testing tool, which allows testing of an assumption applicable to a population**.**

Wilcoxon signed rank test. This is a non-parametric statistical hypothesis test used to compare two related samples, matched samples, or repeated measurements on a single sample to assess whether their population mean ranks differ (i.e. it is a paired difference test). It can be used to determine whether two dependent samples were selected from populations having the same distribution.

Cronbach's Alpha reliability test. This test is conducted in order to measure the internal consistency i.e. the reliability of the measuring instrument (Questionnaire). It is most commonly used when the questionnaire is developed, using multiple Likert scale statements and therefore to determine if the scale is reliable or not.

5. Choosing tools

In seeking to identify the tools required by each of the 7 delivery mechanism Groups within the Science Literacy Models research study, no straightforward division has been possible. For the great majority of delivery mechanisms, few comprehensive or highly illustrative ready-made practical toolkits, best practice definition or models have been published. Nor would it be useful to try to develop a toolkit for each Group because each project or programme needs to design IA in line with the specific circumstances of each intervention. However, it is possible that future IA implementations within science literacy will be able to either emulate or adapt an approach already taken (the published resources organised in this work are a useful guide in this respect), or to construct an approach which meets their specific needs taking all the above research types and design, sampling methods, data collection tools, statistical software examples and statistical methodologies identified from the SLM study, in mind.

Possibly, the main exception to this is the well-established work around **Group 1: Events, Meetings, Performances,** such as:

- *Falk and Storksdieck. Learning science from museums, 2005*, investigates the Contextual Model of Learning to elucidate the complex nature of science learning from museums. http://dx.doi.org/10.1590/S0104-59702005000400007
- *Grand and Sardo. What Works in the Field? Evaluating Informal Science Events, 2016,* describes how the informality of the events and venues in festivals should be reflected in the use of unobtrusive and minimally disruptive evaluation methods. https://doi.org/10.3389/fcomm.2017.00022
- *Jensen and Buckley. Why people attend science festivals: Interests, motivations and self-reported benefits of public engagement with research, 2014,* investigates perspectives of visitors at the Cambridge Science Festival, one of the larger in the UK and provides an example of methodological triangulation. https://doi.org/10.1177%2F0963662512458624
- *Tucker, Bricker, and Huerta. An Approach to Measuring Impact and Effectiveness of Educational Science Exhibits, 2017,* outlines a ten-step mixed-methods approach. https://doi.org/10.4148/1051-0834.1172

In **Group 2: Education and training - including online,** useful guidance exists in:

- *Brinson. Learning outcome achievement in non-traditional (virtual and remote) versus traditional (hands-on) laboratories: A review of the empirical research, 2015.* https://doi.org/10.1016/j.compedu.2015.07.003
- *Fetaji and Fetaji. E-Learning Indicators: A Multidimensional Model For Planning Developing And Evaluating E-Learning Software Solutions, 2009*, Investigates possible approaches to systematic planning, development and evaluation of e-learning initiatives and their corresponding e-learning projects. The result of this work is a multidimensional model of e-learning Indicators that are defined as the important concepts and factors. http://www.ejel.org/volume7/issue1/p1

- *Hahs-Vaughn et al. Statistical Literacy as a Function of Online Versus Hybrid Course Delivery Format for an Introductory Graduate Statistics Course, 2017,* study examines statistical literacy of graduate students using the Comprehensive Assessment of Outcomes in Statistics (CAOS) tool across two increasingly popular delivery formats—hybrid and online. https://doi.org/10.1080/10691898.2017.1370363
- *Lahti et al. Impact of e-learning on nurses' and student nurse's knowledge, skills, and satisfaction: A systematic review and meta-analysis, 2014,* covers a systematic review and meta-analysis of 11 randomized controlled trials (RCT), and assesses their methodological quality. https://doi.org/10.1016/j.ijnurstu.2012.12.017
- *Ruggeri, Farrington, and Brayne. A Global Model for Effective Use and Evaluation of e-Learning in Health, 2013.* Article outlining the key issues for developing successful models for analysing e-health learning. https://doi.org/10.1089/tmj.2012.0175
- *Surpless, Bushey, and Halx Developing Scientific Literacy in Introductory Laboratory Courses: A Model for Course Design and Assessment, 2014.* To assess student learning, the authors used in-class observations, student–instructor discussions, pre- and post- learning questionnaires, prelaboratory quizzes, course activities completed during class time, modified postactivity reflection questions, practical examinations, and a final examination. https://doi.org/10.5408/13-073.1
- *Wu, Chen, and Chen. An Adaptive e-Learning System for Enhancing Learning Performance: Based on Dynamic Scaffolding Theory,* 2017. The study aims to fill the void in the existing literature by building an adaptive eLearning system with self-assessment rubrics based on the dynamic scaffolding theory in response to different student needs. https://doi.org/10.12973/ejmste/81061

Group 3: Traditional publishing and journalism - print and broadcast encompasses a wide variety of mechanisms with many variations on basic methodology, including:

- Numerous relatively concise examples of quasi-experimental approaches (there are several hundred relatively similar projects in the literature) such as: http://ijmess.com/volumes/volume-VI-2017/issue-IV-12-2017/full-1.pdf
- More complex quasi- experimental procedures with greater emphasis on randomisation such as https://www.iiste.org/Journals/index.php/JBAH/article/viewFile/5677/5790
- *Karatsolis. Assessing visual literacy: The case of poster presentations,* 2012, presents three distinct models of assessing poster presentations within the context of an introductory professional communication course as part of a science-based curriculum. The model which was based on a rhetorical approach, where the emphasis was placed on the audience, purpose and context of presentation for the poster, yielded the best results of the three models, both in terms of meeting desired learning outcomes and in terms of student satisfaction. https://doi.org/10.1109/IPCC.2012.6408627

Group 4: Activities and services sometimes feature **r**elatively simple interview-based approaches such as this one for public library makerspaces

- *Gahagan. Evaluating Makerspaces: Exploring methods used to assess the outcomes of public library makerspaces, 2016.* https://researcharchive.vuw.ac.nz/xmlui/bitstream/handle/10063/5193/report.pdf?sequence=1

Group 5: Online interactions includes innovative approaches to assessment of newer digital delivery mechanisms

- *Chen Q, Carbone ET. Functionality, Implementation, Impact, and the Role of Health Literacy in Mobile Phone Apps for Gestational Diabetes: Scoping Review, 2017.* (mobile apps in health literacy) https://doi.org/10.2196/diabetes.8045

- *Maredia et al. Can mobile phone-based animated videos induce learning and technology adoption among low-literate farmers? A field experiment in Burkina Faso, 2018.* Experimental design (4 treatment arms, randomised control trial) and pre-post treatment data). https://doi.org/10.1080/02681102.2017.1312245
- *Prakash et al. Short-duration podcasts as a supplementary learning tool: perceptions of medical students and impact on assessment performance, 2017.* Podcasts for medical students. https://doi.org/10.1186/s12909-017-1001-5.
- Sin et al. Impacts of Social Media Usage on the Outcomes of Students' Everyday Life Information Seeking, 2014. Social media and student information seeking https://doi.org/10.1002/meet.2014.14505101119.
- Chen, Tung-Liang, and Liwen Chen. Utilizing Wikis and a LINE Messaging App in Flipped Classroom, 2017. https://doi.org/10.12973/ejmste/81548.

Group 6: Multiliteracies/Multimodalities

- *Buckley-Walker et al. Evaluating the validity of the online multiliteracy assessment tool, 2017*. Study aims to assess the validity of the Online Multiliteracy (o-Mlit) Assessment for students in Years 5 and 6. The Online Multiliteracy Assessment measures students' abilities in making and creating meaning, using a variety of different modes of communication, such as text, audio and video.https://doi.org/10.1177%2F0004944117729056
- *Jacobs. Designing Assessments: A Multiliteracies Approach, 2013*, contains an overview of multiliteracies approaches to assessment, and explores suggestions of what a multiliteracies approach to assessment can look like in an instructional setting already overly filled with traditional literacy assessments. https://doi.org/10.1002/JAAL.189

Group 7: Citizen science where an econometric model is used to analyse how participation in citizen science activity can contribute to Science Literacy.

- *Masters. Science Learning via Participation in Online Citizen Science, 2016.* https://arxiv.org/ftp/arxiv/papers/1601/1601.05973.pdf