

Beginner's Guide to Measuring Educational Outcomes in CEhp

The Origin of the Series

Pesha Rubinstein, MPH, CCMEP, American Medical Informatics Association

No report in recent history was a more urgent call for action to the medical community than the Institute of Medicine's *To Err is Human: Building a Safer Health System*,¹ issued in 1999. That report stated that up to nearly 100,000 people die in the US each year due to preventable medical errors.

In 2001, the IOM produced its recommendations for addressing the problems outlined in *To Err is Human. Crossing the Quality Chasm: A New Health System for the 21st Century*,² included amongst its recommendations the involvement of continuing education professionals in the transition to a health care system committed to substantial improvement in the quality of health care.

By 2007, the Agency for Healthcare Research and Quality published its report on the effectiveness of CME,³ concluding that CME seemed to be a good thing, but that the evidence was weak.

In the years since, our profession—whether we describe it as CME, CE, CEhp, or CPD—has taken to heart the need to measure the impact of our endeavors. One can see the change by comparing the types of presentations at the annual Alliance conference from a decade ago to those that you will find today. Scanning the types of job descriptions that now exist for CE positions demonstrates how the CE enterprise has transitioned from a documentation-centric one to one that has woven measuring educational impact into educational design.

To remain competent, CE professionals must engage in a continuing education process of our own. Today's CE professional needs a more analytical understanding of medical literature to identify true gaps. Today's CE professional must have a better handle on educational assessment to write outcomes-based learning objectives, and

craft educational offerings resulting in measurable impact.

There are open access courses in statistics available to all that can help CE professionals begin to learn these skills. But which ones are the most relevant to what we do? And for some, the mere mention of “statistics” creates a barrier to learning.

Gary Bird, PhD, of the American Academy of Family Physicians, has assembled a team of CE professionals committed to addressing our own educational needs. The series target audience is the CE professional with little or no experience in educational assessment. The content is meant to be approachable and applicable to the CE professional's work endeavors. After participating in the entire series, the CE professional should be better able to:

- Critique peer-reviewed literature to assess its validity and significance
- Incorporate qualitative and quantitative analytical approaches into the design and planning of CE activities for health care professionals

So what was the origin of the series? In 2010, the CME provider I worked for closed its doors, and I took the opportunity to enroll in a Masters in Public Health program. I took my first biostatistics course ever and realized how important the subject was to research and education and truly how little I knew about educational assessment. I concluded I couldn't be the only CE professional with this educational gap, and thought a series on biostatistics would be relevant for Alliance members. However, I didn't feel qualified to write it. At the 2014 Alliance conference I heard Gary present a session, using an innovative TV-interview format. The session was “Data, Data,

Data: Exploring the Limitations of Competence and Performance Level CME Outcomes,” and it was comprehensible and fun.

After last year's Alliance meeting, I approached Gary about authoring a beginner's guide to using statistics to measure CE activity outcomes. He agreed and assembled the team, who as a group identified the main topics. We are happy to move forward with sharing the results of this teamwork.

The authors are keeping in mind all along that this series is for beginners only. We encourage you to stick with the series, use any tools referred to in the articles, and share them with your colleagues.

Stick with the series, use the tools, share them with colleagues, and provide us feedback.

This series is designed to engage members in active dialogue and feedback. In the coming months we will establish a communities location for your input and comment. Look for an update in the March *Almanac* issue.

References

¹ Kohn LT, Corrigan JM, Donaldson MS, eds; Committee on Quality of Health Care in America, Institute of Medicine. *To Err is Human: Building a Safer Health System*. Washington DC: National Academies Press; 2000.

² Committee on Quality of Health Care in America. *Crossing the Quality Chasm: A New Health System for the 21st Century*. Washington, DC: National Academy Press; 2001.

³ Marinopoulos SS, Dorman T, Ratanawongsa N, Wilson LM, Ashar BH, Magaziner JL, Miller RG, Thomas PA, Prokopowicz GP, Qayyum R, Bass EB. Effectiveness of Continuing Medical Education. Evidence Report/Technology Assessment No. 149 (Prepared by the Johns Hopkins Evidence-based Practice Center, under Contract No. 290-02-0018.) AHRQ Publication No. 07-E006. Rockville, MD: Agency for Healthcare Research and Quality. January 2007.

Beginner's Guide to Measuring Educational Outcomes in CEhp

Introduction to the Series

Gary C. Bird, PhD, American Academy of Family Physicians

Outcomes Data: Where We Are and Where We Are Going

The question of “what does success look like?” is paramount in continuing education (CE). A robust needs assessment is required to define the practice gaps across the spectrum of the health professions, which in turn inform learning objectives, the educational modalities that will be used to bridge those gaps and ultimately, development of the metrics that will address the success question.

However, not all metrics are equal, and the quality of the outcomes produced can vary dramatically. As defined by Don Moore et al¹ and shown in Figure 1, education to bridge a single gap can provide outcomes ranging from the number of learners who attended an educational session (level 1), all the way through to the impact on community patient health (level 7). Clearly, the metrics involved and outcomes derived are not the same based on their ease of measurement, cost to obtain, and most importantly for us as educators, our ability to use them to assess the way the education is changing practice. Yet outcomes of at least knowledge change (level 3), are increasingly required of us to prove our activities have value and to teach us more about the educational needs of the learners we serve.

Measurement of outcomes—and proving the value of our activities—involves an ever-increasing burden of data than most of us have been used to seeing.

Furthermore, measurement of outcomes involves an ever-increasing burden of data than most of us have been previously used to seeing. As a general rule of thumb, as the level of outcome increases, so too does the complexity and potential volume of the data produced, thereby the steeper and harder climb we must undertake. Although

“CLIMBING THE OUTCOMES MOUNTAIN.” AS THE LEVEL OF OUTCOME FROM AN ACTIVITY INCREASES, SO TOO DOES THE VALUE OF THE DATA BUT ALSO THE CHALLENGE FOR THE CE PROFESSIONAL TO OBTAIN THE DATA AND EFFECTIVELY USE IT TO IMPROVE THE QUALITY OF THE EDUCATION.

traditionally, CE providers along with expert faculty have been strong at designing, organizing, and executing educational activities—data analysis and statistics have not been part of the core skills of our profession.

Increasing Data in Higher Level Outcomes: Mountain From a Mole Hill

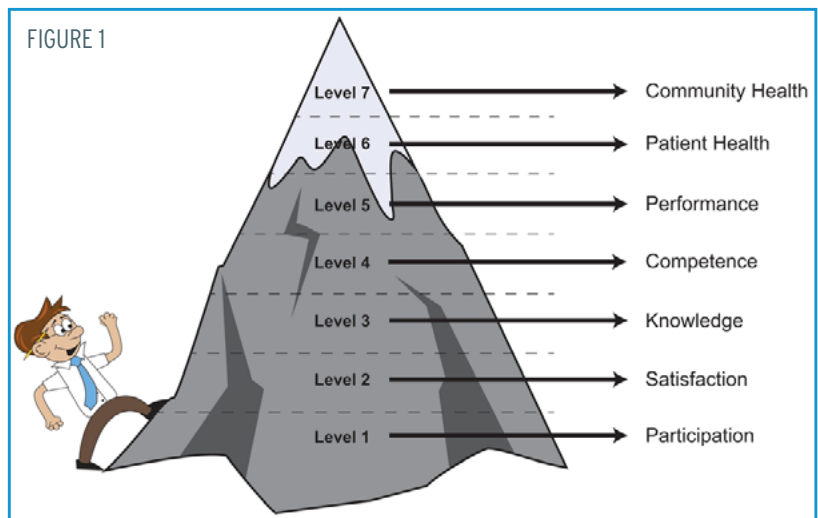
Figure 2 (next page) shows how a “simple” CE activity based on just 100 participants can theoretically yield a very large data set. This small example, which could produce more than 200,000 data points, make the majority of us wistful for the days of simple math problems found on Sesame Street. To this large volume of data, multiple questions can be asked requiring varying complexity of data analysis and manipulation. Analyses can vary from simple assessments of mean, mode, and median, all the way through to complex multivariate analysis and post-hoc testing to assess the effect of the CE activity on community health outcomes.

Why Statistics Matter

47.3% of all statistics are made up on the spot.

—Steven Wright

Fundamentally, statistics provide a mathematical way to describe and characterize data sets and, when necessary, show that two data sets are ‘different’ or ‘not different’ based on defined criteria, as opposed to simply eyeballing the data and guessing. Unfortunately, with the increasing amount of statistical ‘facts’ being bantered about in the current age, the burden and need for critical thinking increases to deci-



Beginner's Guide to Measuring Educational Outcomes in CEhp

FIGURE 2
THE POTENTIAL IMPACT
OF OUTCOMES LEVEL
ON THE AMOUNT OF
DATA PRODUCED IN A
CE ACTIVITY WITH 100
PARTICIPANTS.

Outcome Assessed	Scenario	Data Produced
Patient Health/ Community Health (Level 6-7)	100 Physicians attending the activity belong to 3 hospital systems serving a community. From their EHR systems, they each input data on the 5 metrics for 200 patients pre-, and 200 patients post-CE activity (400 patients total)	200,000 Data Points
Performance/ Patient Health (Level 5-6)	100 Physicians attending the activity input data on 5 metrics related to obesity for 20 chart audited patients pre-, and 20 patients post-CE activity (40 patients total)	20,000 Data Points
Knowledge/ Competence (Level 3-4)	100 Physicians attending the activity answer 10 pre-, and 10 post-questions that are based on the course content (20 questions total)	2,000 Data Points
Satisfaction (Level 2)	100 Physicians attending the activity answer 5 questions on their perspectives of the course content and quality	500 Data Points
Participation (Level 1)	100 Physicians attend a CE activity on Obesity	100 Data Points

pher if and which statistics are correct and also meaningful. A good working knowledge of statistics can offer CE professionals the tools to handle the range of data they may face and provide the framework to accurately measure and then relay the success of their CE activity, wherever they are on the mountain. In this matter, the goal of this series is to provide a working knowledge of statistics for data collection, analysis and results interpretation and dissemination.

A good working knowledge of statistics can offer CE professionals the tools to handle the range of data they may face and help accurately measure the success of a CE activity.

What Can You Expect From This Series?

If reading this introduction brings up memories of sitting in an auditorium in Statistics 101, listening to a professor write equation after equation on the board, wondering what exactly it was that was making them so excited and eager to tell you about the importance of p values--don't panic! This series is for those who are not expert statisticians, but rather beginner or intermediate level folks who want to better explore their data and understand its relevance. The intent is not to drown you in statistics theory, but to open up the world of data analysis in an accessible way that will allow you to pick up practical tips and understanding of how you can get better quality data from your educational activities and make the data work for you.

Reference

¹ Moore DE Jr, Green JS, Gallis HA. Achieving desired results and improved outcomes: integrating planning and assessment throughout learning activities. *J Contin Educ Health Prof.* 2009;29(1):1-15

CONTINUED >>



Beginner's Guide to Measuring Educational Outcomes in CEhp

What Will Be Covered In This Series?

Here's what you can expect to appear in the *Almanac* over the next 13 months:

MARCH, 2015 ISSUE:

SOURCES OF DATA IN CE

Simone Karp, RPh, CeCity

MAY, 2015 ISSUE:

HOW TO ASK GOOD EVALUATION QUESTIONS

Sandra Binford, MAEd, CME Outfitters

Erik Brady, PhD, CCMEP, Clinical Care Options

JUNE 2015 ISSUE:

CONCEPTS INVOLVED IN SAMPLING DATA

Melanie D. Bird, PhD, American Academy of Family Physicians

Erik Brady, PhD, CCMEP, Clinical Care Options

JULY 2015 ISSUE:

IMPACT OF SAMPLING AT VARIOUS SET TIME POINTS AFTER AN EDUCATIONAL INTERVENTION

Sandra Binford, MAEd, CME Outfitters

Gary C. Bird, PhD, American Academy of Family Physicians

AUGUST 2015 ISSUE:

BASIC CONCEPTS OF DATA SETS

Melanie D. Bird, PhD, American Academy of Family Physicians

Derek T. Dietze, MA, FACEHP, CCMEP, Improve CME

SEPTEMBER 2015 ISSUE:

DISTRIBUTION AND VARIATION IN DATA SETS

Gary C. Bird, PhD, American Academy of Family Physicians

Tanya Horsley, PhD, The Royal College of Physicians and Surgeons of Canada

OCTOBER 2015 ISSUE:

HOW TO ANALYZE YOUR PRE/POST ACTIVITY CHANGE DATA PART 1 (CATEGORICAL DATA SETS)

Erik Brady, PhD, CCMEP, Clinical Care Options

Derek T. Dietze, MA, FACEHP, CCMEP, Improve CME

NOVEMBER 2015 ISSUE:

HOW TO ANALYZE YOUR PRE/POST ACTIVITY CHANGE DATA PART 2 (CONTINUOUS DATA SETS)

Erik Brady, PhD, CCMEP, Clinical Care Options

Derek T. Dietze, MA, FACEHP, CCMEP, Improve CME

FEBRUARY 2016 ISSUE:

UNDERSTANDING THE IMPACT OF DATA AND ITS ANALYSIS AT THE POPULATION LEVEL

Gary C. Bird, PhD, American Academy of Family Physicians

Tanya Horsley, PhD, The Royal College of Physicians and Surgeons of Canada

MARCH 2016 ISSUE:

SUMMARY OF THE SERIES

Gary C. Bird, PhD, American Academy of Family Physicians

Pesha Rubinstein, MPH, CCMEP, American Medical Informatics Association.

Beginner's Guide to Measuring Educational Outcomes in CEhp

Sources of Data in CE

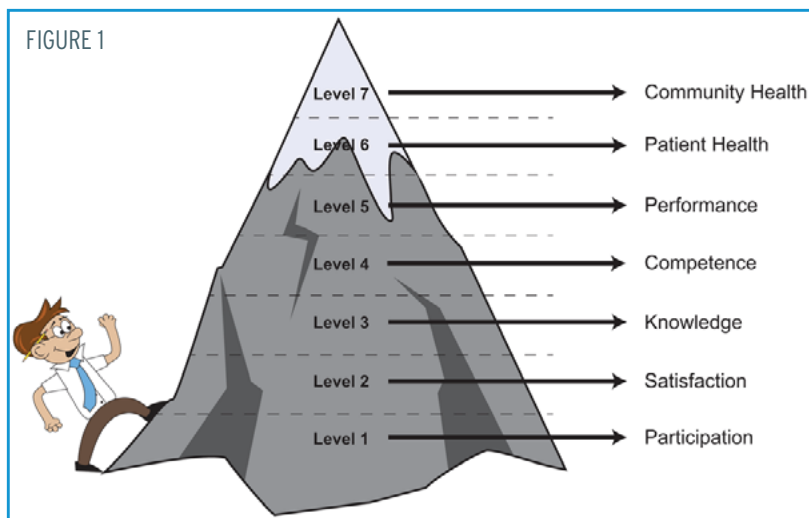
Lloyd Myers, RPh - CECity CEO
Simone Karp, RPh - CECity CBO

Overview of Moore's Model and Sources of Data

Applying Moore's conceptual framework for higher level outcomes assessment in CME leads CME professionals into an increasingly complex world of data, foreign nomenclatures and operational challenges (Figure 1).

At first glance, the upper levels (Levels 5–7) may appear to be overwhelming to many CME providers. CME professionals are generally comfortable with data sources such as attendance records, questionnaires, pre- and post-tests, observations and self-reported results, which are required for levels of Participation, Satisfaction, Knowledge and Competence (Levels 1–4).

However, that same confidence level is not apparent at the higher levels when planning outcomes assessments of Performance (Level 5), Patient Health (Level 6) or Community Health (Level 7), each of which requires the use of one or more sets of clinical, administrative, pharmacy or other type of practice-based data to generate the desired



outcomes evaluation (See Figure 2).¹ Unlike questionnaires and similar instruments, clinical care documents tend to be outside of the control of the CME planner, and therefore present a perceived barrier to participation.

Learning Objectives of this Module

The goal of this module in the Beginner's Guide to Measuring Educational Outcomes in CEhp is to help build confidence through an understanding of the types of data, and sources of data, required to measure performance and achieve higher level outcomes (Levels 5–7) as defined by the Moore model. But, first, it is important to understand how to practically plan activities to assure that you can successfully begin achieving higher levels of outcomes for your organization.

Planning Rules for Performance-based Outcomes

Rule Number 1

Work Backwards from the Measure.

It sounds simple, but we have seen many people try to build their educational plans and strategies based on goals that are unachievable, due to the fact that the data required to measure success is beyond reach, too expensive to acquire, incomplete, fragmented, or due to any number of other reasons. If you start any initiative by first considering what one is attempting to measure, and then working in reverse to

FIGURE 2
FROM MOORE ET AL.¹
REPRINTED WITH
PERMISSION.

Expanded CME Framework	Description	Source of Data
Participation LEVEL 1	The number of physicians and others who participated in the CME activity	Attendance records
Satisfaction LEVEL 2	The degree to which the expectations of the participants about the setting and delivery of the CME activity were met	Questionnaires completed by attendees after a CME activity
Learning: Declarative knowledge LEVEL 3A	The degree to which participants state <i>what</i> the CME activity intended them to know	<i>Objective:</i> Pre- and posttests of knowledge. <i>Subjective:</i> Self-report of knowledge gain
Learning: Procedural knowledge LEVEL 3B	The degree to which participants state <i>how</i> to do what the CME activity intended them to know how to do	<i>Objective:</i> Pre- and posttests of knowledge <i>Subjective:</i> Self-report of knowledge gain
Competence LEVEL 4	The degree to which participants <i>show</i> in an educational setting <i>how</i> to do what the CME activity intended them to be able to do	<i>Objective:</i> Observation in educational setting <i>Subjective:</i> Self-report of competence; intention to change
Performance LEVEL 5	The degree to which participants <i>do</i> what the CME activity intended them to be able to do in their practices	<i>Objective:</i> Observation of performance in patient care setting; patient charts; administrative databases <i>Subjective:</i> self-report of performance
Patient health LEVEL 6	The degree to which the health status of patients improves due to changes in the practice behavior of participants	<i>Objective:</i> Health status measures recorded in patient charts or administrative databases <i>Subjective:</i> Patient self-report of health status
Community health LEVEL 7	The degree to which the health status of a community of patients changes due to changes in the practice behavior of participants	<i>Objective:</i> Epidemiological data and reports <i>Subjective:</i> Community self-report

make sure that all of the data elements are reasonably within reach, the chances of ultimate success rise significantly.

This notion of “working backwards from the measure” is supported by Moore et al² who advised that, when planning, one should “start with the end in mind.” In CME planning, Moore advised beginning with Level 7 outcomes, and then traveling backwards through each of the various levels to better understand where to begin planning activities for providers, based on identified gaps in performance or knowledge. Here, we recommend considering a similar approach based on the desired measures of success at each level, working backwards until you recognize alignment between your gap-analysis, the desired outcomes level, and realistic data sources that can sufficiently power your desired measures.

Rule Number 2

You Can't Improve What You Can't Measure.

In Avedis Donabedian's landmark 1966 paper,³ the founder of modern healthcare quality and outcomes research proposed the Donabedian Model. This conceptual model defines a framework for examining health services and evaluating quality of care,⁴ which includes three categories of measures: structure, process and outcomes.⁵

Although the Moore model is very helpful for planning at the macro level, in order to understand the data source required, one must also consider the type(s) of measures that are to be included within each particular targeted Moore level. The Donabedian Model provides a highly useful way to connect the Moore levels, with the type(s) of targeted measures, and in turn the related data sources required. It also can serve as a framework for analyzing other characteristics germane to data sources that also must be considered, such as data latency (how old is the data), data cadence (how often can I access the data), data refresh rate (how often is the data updated), duration of access to the data, and other criteria that are beyond the scope of this module.

Where Does the Data Come From?

Regardless of the level of outcomes and related type of measures being targeted, access to healthcare data is required to achieve success. Understanding the “data source” required for each measure however, can be confusing due to the myriad of data classification systems in place.

However, one straightforward approach is to use the NQF (National Quality Forum) data source model, which is an integral part of the standard measure specification template used by NQF to define endorsed measures. This data source model is also embedded within the NQF Quality Positioning System™.⁶ By using this data source model as part of your CME planning process, mapping measures to data sources will be simplified.

The following are the NQF defined data sources and a description of where they may be most useful to you in your CME planning process:

- *Administrative Claims*—Administrative claims data, or “claims data,” typically result when healthcare services are utilized and providers submit their

claims for reimbursement. These data can be highly valuable as they include patient demographic information, diagnosis, procedures, provider of care, amount billed and reimbursed for services, and dates of service. A variety of structure and process measures can be calculated based on claims data. The greatest limitation of claims data is that it does not include physiological data elements, such as blood pressure or lab values, and therefore its use in outcomes measures is self-limiting.

- *Paper Medical Records*—The abstraction of data elements from the patient paper-based medical record can be one of the most accurate methods for obtaining clinical data for measuring performance. This data source will provide most of the data needed to power process and outcomes measures. Acquiring data from the patient paper medical record however, is laborious and costly, which makes it difficult to use for upper level outcomes initiatives (Moore Level 6 or 7), and, in general, for any large scale study.
- *Electronic Clinical Data*—Data from electronic data sources, such as EHRs and Clinical Data Registries, hold the most promise as a cost-effective data source for enabling outcomes assessment across all high levels (Moore Level 5–7). The data in these systems are capable of powering structure, process and outcomes measures. Specific to EHRs, the limitations to date have been the limited interoperability available for extracting data, and the lack of complete, or codified data in the EHR. Clinical data registries when used alone, or in combination with EHRs, may provide a more accurate data source to enable large scale outcomes assessment of performance, patient and community health.
 - » Electronic Clinical Data Sources Identified by NQF include: Electronic Health Records (EHRs), Imaging/Diagnostic Studies, Laboratory Systems, Pharmacy Systems, and Clinical Data Registries
- *Healthcare Provider Surveys*—Data from patient responses to healthcare provider surveys have become a permanent fixture in quality measurement and value-based payment programs. These surveys, such as the Centers for Medicare and Medicaid

NQF's online Quality Positioning Systems™ (www.qualityforum.org/QPS/) tool can be very helpful when designing programs that require linking performance measures and data sources.

Services (CMS) HCAHPS (Hospital Consumer Assessment of Healthcare Providers and Systems) patient experience of care survey, can offer a great source of subjective data from the view of the patient, but are limited to the data points and measures prescribed by the survey owner.

- *Management Data*—Practice management system


data, or “PMS” data, mirror “claims” data, but from the provider, rather than the payor, perspective. PMS data result when healthcare services are utilized and include information provided to payors as part of their claims for reimbursement. Similar to claims data, PMS data can be highly valuable as they include patient demographic information, diagnosis, procedures, provider of care, amount billed for services, and dates of service. In addition, the PMS may be able to identify denominator data of all patients across a provider’s practice, which may decrease the burden by limiting data collection needs to only the numerator. An example is in the case of a diabetes measure, where the denominator can be identified by the PMS, and only the HbA1C lab value is abstracted from the patient record.

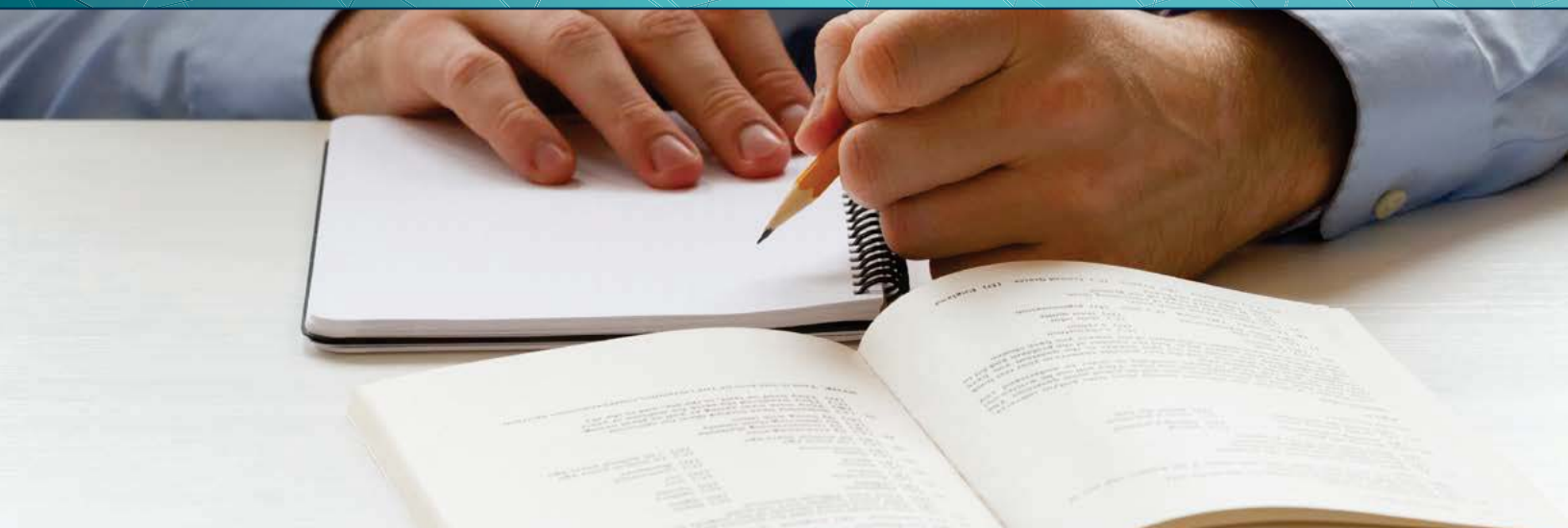
- *Patient Reported Data/Surveys*—Data from patient responses to surveys, including patient reported outcomes surveys, are an emerging data source that hold a great deal of promise for reaching high level outcomes assessment. These surveys tend to be less structured than standardized healthcare provider surveys (e.g. HCAHPS), and may offer a great source of subjective data from the view of the patient.

Case Study

In 2013 the American College of Physicians (ACP) and CE-City developed and launched a quality improvement (QI) clinical data registry (CDR), named the Genesis Registry™ (Genesis). Genesis was conceived based on a needs assessment, which took into account practice-based performance gaps in internal medicine, as well as market-based provider needs related to the shift from service to value. The goals for the registry were set high, and included achieving Performance (Moore Level 5), Patient Health (Moore Level 6) and Community Health (Moore Level 7) outcomes, with support for process and outcomes measures at each level. Using the analysis methods described in this article, the parties realized that this would require continuous data acquisition directly from the practice EHRs. To minimize the burden on providers and EHR vendors, Genesis was designed using electronic enabled measures (eMeasures) with support for standard EHR file formats. Genesis also includes CME (e.g. ACP Smart Medicine™) and other interventions linked to relevant measures, to guide knowledge and performance improvement. Today, the Genesis Registry supports over 5,000 providers and includes over 6 million patients. Continuous performance reports are being generated at the practice performance (Moore Level 5) and patient (Moore Level 6) levels. CME knowledge assessments (Moore Level 3) are also being collected. Additional community outcomes analysis is planned (Moore Level 7), as well as the future inclusion of other measures and data sources.

References

- ¹ Moore DE Jr, Green JS, Gallis HA. Achieving desired results and improved outcomes: integrating planning and assessment throughout learning activities. *J Contin Educ Health Prof.* 2009;29 (1):1-15.
- ² Moore DE Jr, Green JS, Gallis HA. Achieving desired results and improved outcomes: integrating planning and assessment throughout learning activities. *J Contin Educ Health Prof.* 2009; 29 (1):6.
- ³ Donabedian A. The quality of care. How can it be assessed? *JAMA.* 1988 Sep 23-30;260(12):1743-8.
- ⁴ McDonald KM, Sundaram V, Bravata DM, et al. (2007). Closing the Quality Gap: A Critical Analysis of Quality Improvement Strategies (Vol. 7: Care Coordination). Rockville (MD): Agency for Healthcare Research and Quality (US); 2007 Jun
- ⁵ Donabedian A. The quality of care. How can it be assessed? *JAMA.* 1988 Sep 23-30;260(12):1743-8.
- ⁶ NQF Quality Positioning Systems™ (www.qualityforum.org/QPS/) 



How to Write Sound Educational Outcomes Questions: A Focus on Knowledge and Competence Assessments

By Erik D. Brady, PhD, CHCP, Director of Analytics, Reporting, and Outcomes at Clinical Care Options, LLC; and Sandra Haas Binford, MAEd, Senior Educational Designer and Research Analyst at CME Outfitters, LLC

While there are many available resources on authoring well-structured question items in tests and surveys,¹⁻⁷ few articles contemplate the specific needs of the question item-writer in continuing education in the health professions (CEhp). Many principles in item-design translate to the CEhp space, but some specific considerations affect decision making in writing items for CEhp, predominantly due to the need to measure the various changes in knowledge, competence and practice-based performance achieved by clinical learners through their participation in an educational intervention.

The focus of this article, the first in a three-part series, is to provide guidance to CEhp professionals on best practices in question item-writing, an element that is foundational for most of the remaining articles in this series. While many assessment types exist, this article focuses on items that have a defined set of answer options (i.e., close-ended items, such as multiple choice questions and rating scales). Among these close-ended items, assessments can be further categorized into those that a grader observes (correct/incorrect, optimal/suboptimal, etc.) and those that a learner self-reports upon reflection or other self-assessment.

Open-ended questions can and certainly are used in the context of continuing education assessment, and they typically gather valuable qualitative data, such as learners' comments, reflections, questions for future education and ideas for implementation of clinical evidence to routine practice, rather than quantitative data. Analysis of qualitative data is an important part of the overall medical education assessment toolbox; however, for simplicity in this article series, we will focus on item-writing that gives rise to quantitative data sets, which in turn allow for the quantitative analysis methods that will be covered in the articles that follow.

Finally, readers who are familiar with the outcomes levels defined by Moore, Green and Gallis⁸ will recognize our focus on Levels 3, 4 and 5 (commonly known as knowledge, competence and performance, respectively). We particularly emphasize basic knowledge assessment and competence measurement through case scenario testing, attitudinal indicators and confidence (a measure of self-efficacy). Because all competence types form the glue that links clinicians' knowledge and skills acquisition to routine performance in care practices, we can trace lower-level assessments to planners' goals to change clinician performance through participation in educational content

(Level 5). The assessment of health outcomes measurement among patients of participating clinicians is beyond the scope of this article. **Table 1** categorizes the assessment types that lend themselves to quantitative analysis, highlighting those that we include in the item-writing principles below.

The Assessment Goal

The central goal of any assessment question item should be to accurately measure the learner’s current status. In the case of a knowledge-based assessment, the goal is to author an item that accurately measures the learner’s current knowledge. The goal of a competence-based assessment is to have an item that accurately measures the learner’s intent (e.g., commit to implementing a practice change or making a clinician-appropriate decision in a realistic scenario).

Definition of an Assessment Item

In a discussion of question items, it’s important to review the terminology used to describe a question item. **Figure 1** shows a general example using a multiple-choice framework. The question stem refers to the question itself and may include text that defines a clinical scenario or otherwise describes a condition that the learner needs in order to appropriately consider how to respond to the item. The answer options are the set of responses from which learners may select. For multiple-choice items, options include the key, the correct answer option, and a set of distractors, or incorrect answer options. All of these components together form a test item. For our purposes, test item and assessment item will be used interchangeably throughout this article.

Multiple-Choice Items: The Building Blocks of Robust Assessment

In the case of multiple-choice items, specific technical requirements should be met for each item. At a basic level, an assessment item should conform to the specifications shown in **Table 2**.¹⁻²

A common error in assessment item writing is the construction of assessment items that focus on a minor or trivial data point found in the content. This practice is particularly common when assessment items are written from finished content or when minimum scores needed for learners to request educational credit dictate the number of items on a tool, causing planners to select trivial points in desperation to hit a particular quota.

Because assessment items are optimally designed to assess how well a learning objective has been met, aligning a

Table 1. Types of CEhp Assessment Items: Examples of use of question items to give rise to observed or self-reported measures for Outcomes Levels 3-5

Outcomes Level	Observed measure	Self-reported measure
Level 3 (knowledge)	Multiple choice items designed to assess knowledge change	Rating scale items to assess desired knowledge change Rating scale items to assess degree of knowledge change
Level 4 (competence)	Multiple choice items designed to assess intent to practice	Rating scale items designed to assess planned practice change, confidence change or planned frequency of practice strategy use
Level 5 (performance)		Rating scale items designed to assess incorporation of planned practice changes

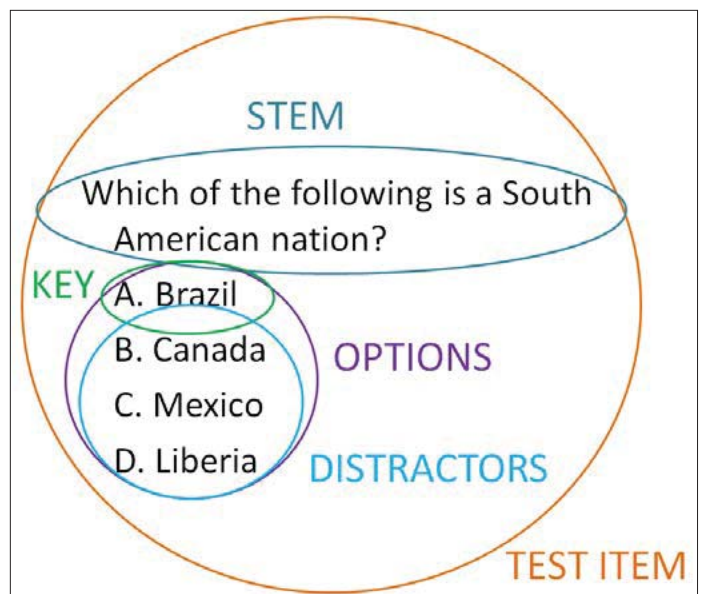


Figure 1. Defining Components of a Multiple-Choice Item.

learning objective with an assessment item should ensure that your assessment items are focused on the key points of the activity content and that activity content consistent-

ly supports learners’ achievement of the educational and performance objectives. This allows for better assessment of the activity content segments and also offers a platform for ongoing needs assessment.

The Question Stem

Question stems should be clear and free of excess language and detail. If too much detail is included in the stem, learners may focus on minutiae that do not directly inform their response to the assessment item. On the other end of the spectrum, the stem should be specific and focused. If not enough detail is included, learners will not be able to determine which of the options is the key, even in cases where the learner is well informed on the topic. When this occurs, the overall goal of the assessment item has not been met as an accurate assessment has not been achieved.

A best practice that is often challenging in medical education is avoiding negative phrasing in the question stem. There are two major issues with the use of negative phrasing, or using words such as “not” or “except,” in a stem. First, studies⁹ have shown that negation makes items significantly more difficult for the learners to comprehend. It is not unusual for learners to have to re-read an assessment item that includes negative phrasing in order to select the most appropriate answer, hindering the learner’s current status on the topic of the assessment item. Second, assessment authors can find it difficult to avoid negation when incorporating the need for a key to be reference supported. Often, when constructing an item that incorporates negative phrasing, the distractors are reinforced by data, while the key often lacks that same level of literature support. For both reasons, the best practice is to avoid the use of negative phrasing in assessment items.

The Answer Options

Many of the same principles that govern best practices for stems apply here as well, for example options should be clear, free of excess detail, specific and focused.⁵ The options do have some additional criteria that must be considered to enable the construction of a strong assessment item. Options must be relevant and plausible to the situation described in the stem. Failure to achieve this invariably leads to easy elimination of some of the answer options by savvy test subjects, which will likely result in a higher percentage of correct scores than is accurate.

The options should be similar to one another in both detail and length. An answer option that has significantly more detail than other answer options or is significantly

Table 2. The Components of Strong Question Items.

Characteristic	Specification
Difficulty of Item	The item should assess minimum competency; it should not challenge the most knowledgeable learners.
Target Audience Appropriate	The item should be free of all types of bias (e.g., age, ethnicity, gender, race, region, and religious and commercial-related support).
Free of Bias	Does something belong here?
Reference Supported	The key, in particular, should be supported by research or other reputable data.
Relevance	The item should focus on a major teaching point, or aligned to a learning objective, and it should not assess a minor or trivial data point found in the informing content.
Clarity/Brevity	The item should avoid use of excessive language and detail. It should also have correct grammar and punctuation.

longer than the other answer options is often a cue to the respondent that helps identify the key.

Likewise, the use of absolutes, such as “always” or “never” are cues to the learner to eliminate those options from consideration. It is also a best practice to avoid the use of “all” or “none of the above” in assessment item writing.⁶ The use of these as either keys or distractors routinely leads to an inaccurate assessment of the learner’s current status. Further, when “all of the above” is used as the key in an assessment item, all of the options are correct to a degree, and there are no true distractors in the item.

Options should avoid using wording from the stem, a best practice that is particularly true for the key. When the best answer picks up wording from the stem, it serves to cue the learner to the key.⁶ The crucial point of the distractors is to be clearly suboptimal. In medical education, it can be difficult to develop clearly suboptimal yet plausible distractors. Analysis of the data consistently becomes more challenging when an option that was intended to serve as a distractor is later revealed to be an optimal answer.

The best practice is to include no fewer than three distractors, with four distractors as the preferable number.¹ In general, try to maintain a situation in which the respondent has no more than a 25 percent chance of guessing the correct answer. One technique is to include an answer option that allows the respondent to clearly indicate that they do not know the optimal answer, or an unsure option. When used, it does offer a powerful assessment of educational need, as the respondent has unequivocally stated that they are in need of education on the topic. As there is no specific upper limit on the number of distractors in an assessment item, except in consideration of the need for item brevity, inclusion of an unsure option offers a straightforward way to enhance the data set. It is important that item writers do not consider unsure as one of the main distractors, as this is clearly not a plausible one.

Optimizing Multiple-Choice Questions Across the Assessment Tool

A handful of additional considerations must be made regarding the design and use of a set of assessment items for a particular activity. This set of questions is often referred to as the assessment tool. One of the main concerns in constructing an assessment tool is the need to ensure that the situation sets for each item are independent of one another. In addition, it is necessary to avoid providing a sequence of questions in which the correct answer to the first item is described in the question stem for any subsequent item. However, this is not to argue that a single case scenario cannot give rise to multiple, independent assessment items, but it does demand that subsequent items are not designed to rely upon a specific answer to a previous item.

It is also best when the optimal answer is not consistently in the first, or “A,” position. Deliberately arranging the answer options to ensure the correct/optimal answer is not consistently in the same position across the assessment tool is worth considering.

Guidance on Dichotomous (Two Option) Items

True/False: While technically not ruled out as a best practice in item writing, there are some specific requirements for its use. True/false can be used to assess knowledge and comprehension, but it may fall short of a competence measure, as a learner’s intent cannot be assessed using a true/false item. In addition, respondents often have to guess what the item author intended with the item for cases in which the options are less than completely true or completely false.^{1,10} For these reasons, more and more educational providers are avoiding the use of true/false items in clinical education.

Yes/No: If the goal is to assess the usage of a specific practice strategy, a yes/no item can be used. It is distinct from a true/false item when used to measure a learner’s performance or frequency of practice strategy use. Using yes/no on a knowledge item, however, is typically not distinct from a true/false item.

Rating Scale Items

The use of rating scale items in medical education is very common in assessing Level 2 (satisfaction) outcomes, and there is growing interest in their appropriate use for higher levels of outcomes. Many of the characteristics described in **Table 2** still apply with rating scale items. For example, sound rating scale items should be appropriate for the target audience, clear and focused, and relevant to the major teaching points. They can be used to measure degree of agreement, confidence and frequency of use/planned use, among other things.⁷ While rating scales do not have a best answer like multiple choice items, they are still powerful items that allow for an assessment of intent to practice that can feed into a performance measurement at a later point in time.

In addition to the considerations outlined regarding multiple choice items, a rating scale item needs to be as specific as possible. If too little detail is included in a rating scale item, the respondent is likely to select a higher rating than is accurate. Additionally, it is critical to ensure that the stem represents a single testing point only.

While Likert scales are the most used type of scale applied to rating scale items, many authors have found that semantic differential scales are both more flexible and more reliable. Rating scales should be unbiased, keeping the mid-point of the scale neutral and balanced, meaning the space between each option is equal. Use of Likert scales are challenging specifically on the latter point. It can be difficult to verify that the difference between excellent and very good is the same as the difference between fair and poor. Rating scales should have at least five points, and a body of research suggests that a seven-point scale is better than a five-point scale.¹¹

Assessment items, when incorporated into the content planning process, can be used along with the stated learning objectives to help content developers focus and fine tune the development of the content, providing a frame against which to guide faculty and review draft content. Therefore, to use assessments to guide content inclusions, assessment items must be discussed and authored early and in collaboration with content developers and educational designers, not as an afterthought that takes place towards the close of activity


Table 3. Comparing Rating Scales: Tailoring Assessment Items to CEhp

Example Item	Comments
<p>“All” or “None” of the above</p> <hr/> <p>Under which circumstance(s) should you temporarily suspend bevacizumab?</p> <p>A. Hypertensive crisis 3%</p> <p>B. Hypertensive encephalopathy 1%</p> <p>C. Severe hypertension that is not controlled with medical management 8%</p> <p>D. All of the above 85%</p> <p>E. None of the above 4%</p>	<p>This item is shown with the percentages of those selecting each answer option on the right hand side. This item failed to consider two specific best practices: The use of all of the above and the need for similar levels of detail and length for each answer option. In this case, the optimal answer was “C”, but learners simply could not avoid selecting “D”. The next most selected response was the longest option, “C”</p>
<p>What is the minimum blood flow requirement to maintain normal cerebral function?</p> <p>A. 20mL/100 g/min</p> <p>B. 40 mL/100 g/min</p> <p>C. 50 mL/100 g/min</p> <p>D. 60 mL/100 g/min</p>	<p>This knowledge item meets many of the requirements of a sound item: All options are of equal length and detail and it is clear and brief. A minor suggestion would be to move the “mL/100 g/min” into the question stem, as it is consistent across all of the answer options.</p>
<p>Based on your understanding of the CDC guidelines, which of the following statements best describes the population of patients to whom you should recommend HIV testing?</p> <p>A. All individuals 13-64 years of age, regardless of individual risk factors</p> <p>B. Any patient with ≥ 1 individual risk factor</p> <p>C. Any adult patient who is member of a high-risk group or who is pregnant</p> <p>D. All adult patients if the HIV prevalence in your setting exceeds 2%</p>	<p>This is another reasonable knowledge item, focused on assessing what the learner knows, rather than what the learner intends. The options are of similar length and detail.</p>
<p>A genotype 1 HCV-infected patient has completed three weeks of triple therapy with telaprevir and is experiencing a mild to moderate telaprevir-associated rash. How would you choose to manage this patient’s rash?</p> <p>A. Begin systemic steroids</p> <p>B. Begin topical steroids (optimal)</p> <p>C. Reduce his dose of telaprevir</p> <p>D. Stop the telaprevir and then restart when his rash has resolved</p> <p>E. Unsure</p>	<p>This is an example of a sound competence item. Note the phrasing of the question stem; rather than asking the learner what is best, it instead asks how would you, language that moves the assessment item into the realm of competence. In addition, “unsure” is included as an option to offer the respondent an opportunity to clearly indicate uncertainty.</p>
<p>How confident are you in your response to the last question?</p> <p>A. Extremely confident</p> <p>B. Very confident</p> <p>C. Confident</p> <p>D. Neutral</p> <p>E. Not confident</p>	<p>This is an example of a fairly typical Likert scale item frequently used in CE. The scale is biased, as the designer has offered more positive ratings than negative ratings. It is also unbalanced, as the gap between extremely confident and very confident is not equal to the gap between neutral and not confident.</p>
<p>How confident are you in your ability to select a therapy for patient with ITP?</p> <p>1. Not confident</p> <p>2. Neutral</p> <p>3. Confident</p>	<p>This item offers an unbiased and balanced rating scale, but a three-point scale is likely to result in a minimal measure of change. In addition, the item itself is vague. Most physicians would be likely to evaluate themselves highly at baseline, even when not having high confidence in the patients that may be discussed within the content.</p>
<p>How confident are you in your ability to select second- and third-line therapies for patients with early stage CLL and platelet counts below 30 x 10⁹/L?</p> <p>Not confident ○ ○ ○ ○ ○ ○ ○ Very confident</p> <p>1 2 3 4 5 6 7</p>	<p>This item is strong in that it offers a very specific item that forces the learner to reflect on individual patients like the one described in the stem. In addition, the item is well supported with a seven-point semantic differential scale, which is both balanced and unbiased.</p>

development and on the threshold of activity implementation. The challenges of simultaneous, collaborative content and assessment development are evident in any activity that presents assessment items among content pieces, such as with audience-response questions during live activities or patient simulations with pathways that branch into optimal and less optimal pathways. Issues regarding timing of such questions will be addressed in the next article in this series.

A common fault in the generation of assessment items in clinical education occurs when an item is focused on a decision point that rests solely with the physician members of the care team, but is then used to assess other members of the care team. This is a true challenge in activities that are meant to address the educational needs of care teams, but the reality is that learning objectives and thus their assessments should be role specific. When that paradigm is embraced, it is more likely that assessment items will be specifically crafted for appropriate members of the care team.

When we author assessment items that achieve the goal of accurately measuring the learner's current status, we open a number of possibilities for analysis. If we position the question within an activity prior to informing content, defined as content that allows the learner to better answer the assessment item, we can use the data collected to validate an educational need amongst a group of respondents. If we repeat the question after the informing content, we have a platform from which to discuss ongoing areas of educational need, even for those learners who have participated in our activities. If we do both, we have data allowing an analysis of the educational and even clinical effectiveness of the specific activity in which the question was asked. All of that said, without strong, well-constructed items that can be used as the backbone of our assessment efforts, analysis along these lines becomes more difficult.

The [National Board of Medical Examiners item writing guide](#) is a powerful resource for CEhp professionals who are intent on improving their competence in both the writing and reviewing of multiple choice items. It has countless positive and negative examples that highlight the perils and pitfalls of item writing. 

Coming Up Next

In the next article, Erik D. Brady, PhD, CHCP, and Melanie D. Bird, PhD, will review foundational concepts in sampling of data and will discuss external validity, sampling methods, sampling errors and bias, controls vs. experimental groups, randomization, and some statistics regarding distri-

bution and size of samples. Attention will be given to the positioning of questions to collect baseline and post-activity outcomes for the different question types described in this article to assess changes in knowledge and/or competence in the context of CEhp activities.

Definitions

Close-ended question: A structured question that limits a respondent to a defined list of answer choices

Open-ended question: An unstructured question that encourages respondents to share their knowledge or feelings about a topic in their own words

Quantitative data: Data/information that can be measured or quantified

Qualitative data: Data/information that can be observed but cannot be quantified

REFERENCES

1. National Board of Medical Examiners. Item writing manual, 3rd ed rev. Philadelphia, PA: National Board of Medical Examiners, 2002. www.nbme.org/PDF/ItemWriting_2003/2003IWGwhole.pdf.
2. Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied measurement in education*, 15(3), 309-333.
3. Collins, J. (2006). Writing Multiple-Choice Questions for Continuing Medical Education Activities and Self-Assessment Modules 1. *Radiographics*, 26(2), 543-551.
4. Norcini, J. J., Swanson, D. B., Grosso, L. J., & Webster, G. D. (1985). Reliability, validity and efficiency of multiple choice question and patient management problem item formats in assessment of clinical competence. *Medical education*, 19(3), 238-247.
5. Al-Faris, E. A., Alorainy, I. A., Abdel-Hameed, A. A., & Al-Rukban, M. O. (2010). A practical discussion to avoid common pitfalls when constructing multiple choice questions items. *Journal of Family and Community Medicine*, 17(2), 96.
6. Harasym, P. H., Leong, E. J., Violato, C., Brant, R., & Lorscheider, F. L. (1998). Cuing effect of "all of the above" on the reliability and validity of multiple-choice test items. *Evaluation & the health professions*, 21(1), 120-133.
7. Krosnick, J. A., Judd, C. M., & Wittenbrink, B. (2005). The measurement of attitudes. *The handbook of attitudes*, 21-76.
8. Moore, D. E., Green, J. S., & Gallis, H. A. (2009). Achieving desired results and improved outcomes: integrating planning and assessment throughout learning activities. *Journal of Continuing Education in the Health Professions*, 29(1), 1-15. Tamir P. *Studies in Educational Evaluation*. 2993; 19, 311-325.
9. Grosse, M. E., & Wright, B. D. (1985). Validity and reliability of true-false tests. *Educational and Psychological Measurement*, 45(1), 1-13.
10. Colman, A. M., MORRIS, C. E., & Preston, C. C. (1997). Comparing rating scales of different lengths: Equivalence of scores from 5-point and 7-point scales. *Psychological Reports*, 80(2), 355-362.

Beginner's Guide to Measuring Educational Outcomes in CEhp

Concepts Involved in Sampling Data

By Melanie D. Bird, PhD, American Academy of Family Physicians; Erik D. Brady, PhD, CHCP, Clinical Care Options

Appropriate and accurate research design is essential for successful measurement of outcomes in any educational activity. A major component of the design is deciding who will be included in the study or the sample. Gathering data from all learners and potentially affected patients is time and resource prohibitive, and as discussed in previous articles, the burden of data collection has increased. Therefore, most researchers use discreet and manageable sample populations.

Scope of this Article

The goal of this article is to provide a review on foundational concepts in sampling, discussing sampling terminology, external validity, sampling methods, errors and bias, group formation, and a brief overview of statistics relevant to sample size. It will also focus on collection of baseline and post-activity outcomes in regard to sample composition and positioning of questions to assess change in knowledge or competence.

Importance of Sampling

Sampling is the process of selecting participants from a particular population to represent that population as a whole. You can select a subset of your overall targeted learners, allowing for extrapolation of the results to the entire population of interest. It is an important tool for research in CEhp outcomes, as it is much easier to work with a smaller group instead of a large population and, arguably, saves time and money. It also allows for more control of the study and less risk of human error in data entry and analysis.

In general, sampling requires a statement of who the targeted learners are (e.g., degree type, board certification or specialty, experience) and/or who the learners' patients are (e.g., practice setting, type of patients seen, geography of the practice, etc.). It is important to understand the concept of external validity or the extent that the study results can be generalized back to the population at large. In other words, would all the relevant providers show an increase in knowledge, competence or performance following your educa-

tional activity, or only those directly similar to your sample population?

Probability Sampling

Most studies employ a sampling plan¹⁻² to create the sample population. Examples of these plans can be found in **Table 1**. Probability sampling involves a deliberate and unbiased plan that allows for every sample unit to have an equal chance of being included in the sample.

Simple random sampling would involve selecting participants in such a way that every possible person has an equal chance of being selected, which is the equivalent of drawing names from a hat. This can be a challenging approach in CEhp activities, particularly in instances where a significant percentage of your learners are not members of the activity's targeted audience. Examples include an in-person activity targeted specifically at physicians held in a setting open to many provider types or an online activity that is targeted to health care teams, but does not restrict patients from accessing the content and participating in the outcomes measurement. Both examples allow for non-targeted learners to participate in the outcomes study, making a completely random sampling subject to analysis that leans on data that would otherwise be dropped.

Other categories include systematic sampling, which often involves random, computer-generated numbers used to select participants for the sample population.

Stratified random sampling involves placing participants into mutually exclusive sets, clusters or strata, and then randomly selecting from each set. Examples of strata might include age, sex, practice setting, geographic parameters, etc. By ensuring randomness into the sample selection, we can limit sampling error and subsequent bias in our data and increase the external validity of the study. Random sampling is used in CEhp programs, not only to limit bias but for two other tangible reasons: It requires the least amount of forethought in the design of

Table 1. Examples of Sampling Methods for CE Activities.

Random Sampling	Simple	Participant names are placed in a pool and then are selected one at a time at random to receive a survey following a CE activity.
	Systematic	Participants are assigned a computer-generated number at time of enrollment in a CE activity. Those with particular numbers are sent a survey.
	Stratified	Participants attending a CE activity on atrial fibrillation management are divided into specialty areas: family medicine, cardiology, gerontology, surgeons. Participants from each group (strata) are picked at random to be in the sample.
Nonrandom Sampling	Convenience	A portion of participants in a CE activity are asked to complete a survey based on proximity to the host of the activity.
	Consecutive	All participants at the CE activity complete the survey, but only target learners are included in the sample.
	Snowballing	Participants in a CE activity complete a survey and submit additional names to be contacted for inclusion in the sample.

the outcomes tool, and it allows the analyst to report the highest participation possible in the outcomes study.

Nonprobability sampling

In cases where time, money, or other issues are constraining, investigators may use nonprobability (nonrandom) sampling. In these cases sample units are not selected randomly but are selected based on accessibility or judgment of the researcher. Nonrandom sampling methods are less stringent and widely used; however, there is a greater chance of bias in the sample, decreasing the external validity.

A method for nonrandom sampling includes convenience sampling, which uses easily accessible subjects. If all accessible subjects are included in the group, we call it consecutive sampling. Consecutive sampling is often used in CEhp outcomes analysis. This method allows the analyst to efficiently eliminate non-target audience members, like non-health care providers or specific provider types, when the activity measured is not intended to address their educational needs, while maximizing the number of outcomes participants. Other methods include quota sampling, in which individuals are included in equal numbers in each group based on a specific trait (age, sex, type of practice) or snowball sampling, where recruited subjects are asked to identify others to include in the sample.

Sampling Errors

Two types of error can result from using a sample population: sampling error and non-sampling error. Sampling

“The lower the standard deviation and the larger the sample size, the smaller the sample error becomes.”

error, also called random error, results from differences in the sample compared to the population of interest. For example, even with a random sample, we might end up with too many providers from one geographical region, practice type or specialty. Sampling error is random and out of our control, but can be limited through increased sample size. Sampling error refers to the level of precision and can be expressed in percentage points. For example, if sampling error is low and our level of precision is ± 5 percent then we can expect our results to fall within that range.

Non-sampling error results from a systematic error that can lead to bias in the study. Usually, this error occurs due to mistakes in data entry or acquisition or inappropriate sampling methodology from poor planning and inattention to detail. Non-sampling errors can result from three major areas: errors in data acquisition, non-response errors and selection bias. Errors in data acquisition occur when the recording of responses is incorrect,

due to mistakes made from transcription of primary information, equipment error or faults, inaccurate responses resulting from incorrectly written or ambiguous questions (responder misinterpretation) and more. Non-response error or bias occurs when responses are not obtained from some members of the sample. This type of error results in either a substantially smaller sample size that may no longer be representative of the population, or if the responders' answers are extrapolated to the non-responders, investigators may reach an incorrect conclusion. For example, we may incur non-responder bias to a survey on satisfaction for examination preparation. Participants who scored well may be more motivated to respond than those that did not, thereby skewing the results. Similar to non-response bias, selection bias occurs when some members of a target population cannot be selected for inclusion in the sample. Increasing the sample size will not alleviate non-sampling errors.

Sample Size

Some statistical knowledge is needed in order to understand the importance of sample size.¹⁻⁴ The sample size is dependent on several parameters used in inferential statistics related to the data collected. As outlined in the first three articles of this series, the type and amount of data will vary depending on the outcomes being measured and the number and types of questions used for the assessment. Each response is a data unit, and for each data unit a sampling statistic can be calculated (mean, median, mode). To translate the sampling statistic back to the population of interest, we need to understand the distribution of our sample. The sampling distribution is the spread, or possible values of a statistic, across an infinite number of samples and resembles a bell shaped curve when graphed,¹ as shown in **Figure 1**.

The statistic or parameter observed represents just one of infinite possibilities. The spread of scores around the parameter for our population is called the standard deviation (often abbreviated to SD, or denoted by the Greek letter σ). The spread of scores across the sampling distribution is the standard error (sampling error, or SE).¹ The standard error is calculated using the standard deviation and sample size. The lower the standard deviation and the larger the sample size, the smaller the sample error becomes. The sample size needed for measurement of outcomes for a particular CE activity is dependent on the amount of acceptable error related to how big a difference you would like to find. A small difference would require a large sample size. If you are looking for larger differences, then a smaller sample will be sufficient. **Table 2** outlines how different levels of error require varying num-

Table 2. Sample size required varies on population and precision level (level of error).

Population size	5%	10%
10	10	n/a
50	44	n/a
100	81	51
500	222	83
1,000	286	91
2,000	323	92
10,000	385	99
100,000	398	100

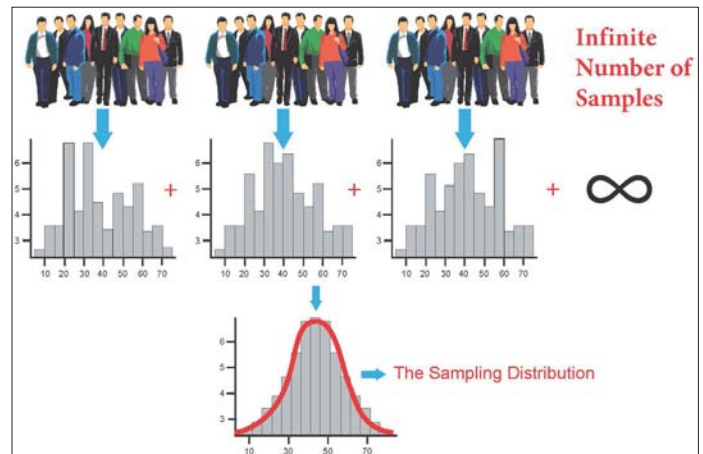


Figure 1: Sampling Distribution: mean, mode, median from an infinite number of samples.

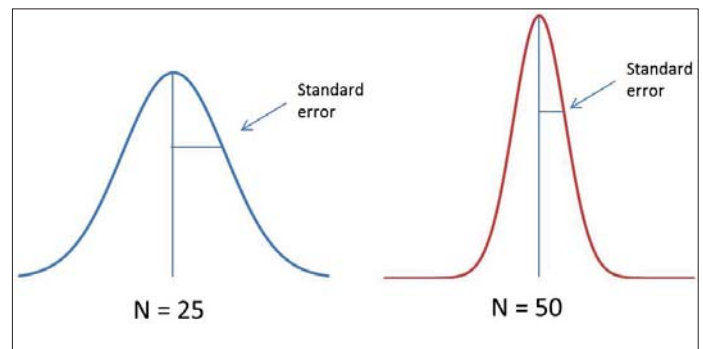


Figure 2: Increased sample size results in a narrower sample distribution and decreased standard error.

bers of participants or respondents.³ For more precise outcome measurement, the error level should be lower, requiring a larger sample size. As shown in **Figure 2**, by increasing the sample size from 25 to 50 participants, the standard error becomes smaller. For a visual explanation of the impact of sample size on sampling error, see the first three minutes of R. Backman’s video “**Sampling Error and Sample Size**”

As discussed in Articles 1 and 2 of this series, in outcome levels 3-7, changes in knowledge, performance and patient/community health are assessed. In order to demonstrate change following an intervention, researchers must create the appropriate groups within their sample: a control group and an experimental group. The experimental group receives the intervention (new educational activity) and the control group receives the standard educational activity or no education depending on the question being evaluated. In order to know how many participants should be included in each group, power analyses can be performed to provide guidance. The precision rate (sampling error), the confidence intervals and the degree of variability all impact the sample size, as seen in **Figure 3**.⁴ The precision rate and confidence intervals reflect how sure one can be of the mean result where the degree of variability depends on the heterogeneity of the sample. The more diverse the sample is, the larger it will need to be to account for the variability and the more confident we are in the result. The actual calculation of sample size can be done using any number of websites and software, such as www.VassarStats.net. After knowing the size of the control and intervention groups, you must divide sample members between them.

Research Design and Sampling

To assess a change in knowledge, competence or performance, multiple research designs with varying number and composition of groups are available. Designs may include one measurement at the end of an activity, or intervention (posttest only), or a pretest and posttest, and may include repeated measures or tests over a specified duration. Activities that use only a single measurement post activity are particularly subject to the considerations previously described, outlining the need for a demographically matched control for the learner set involved in the outcomes study.

A more simple research design involves each participant serving as their own control, thus receiving the pretest, the educational intervention and then the posttest. This design is easy to set up and requires a more limited

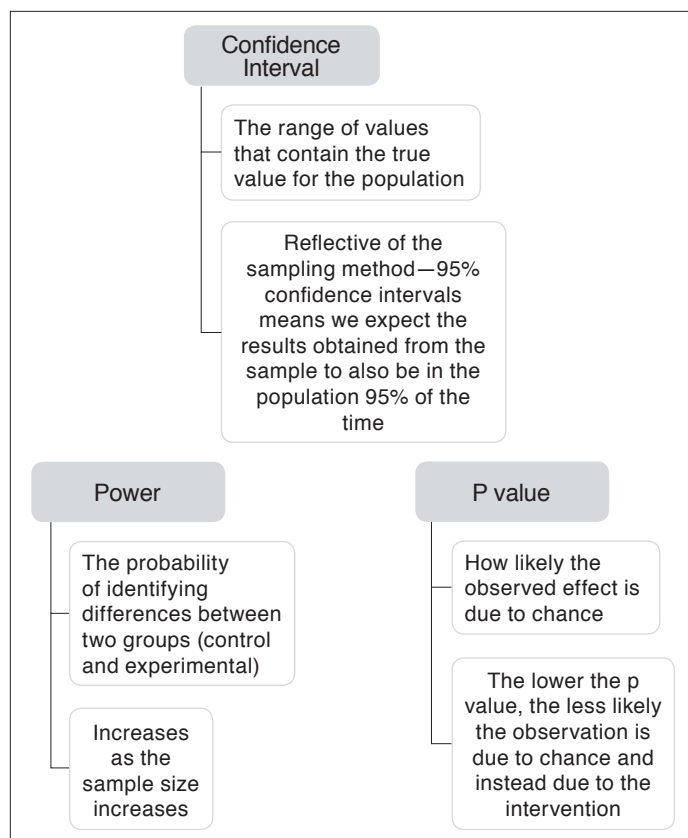


Figure 3: Factors affecting sample size include confidence interval, p value and power of a study.

sample size; however, it can lead to confounding variables with participants performing poorly on the posttest due to fatigue or performing better than expected due to practice. Even so, this design is likely the most typical design found in the CEhp space as the demographic matching of different sample populations is inherently more difficult to manage than item fatigue.

Another simple study set up is the two group posttest only design. In this case, sampling is performed to create a control group and experimental group. Both groups are given a test following the intervention (or control) and the results are compared. The issue with this design is that we cannot be sure if the two groups were similar at baseline without a pretest, which impacts the internal validity of the study.

To determine baseline, both the control and intervention group receive a pretest to determine baseline knowledge and then both groups take a posttest following a specified interval. During the interval, only the experimental group will receive the intervention. A benefit of the pretest-posttest design is that we increase internal validity be-

cause we know the two samples are similar at baseline and can effectively measure a change due to our intervention. A concern for this design is a potential loss of external validity as the pretest may influence the results. For example, participants in the control group previously unconcerned with a topic may do some self-study or get outside information leading to an increase in posttest scores similar to the experimental group. This may hinder our ability to generalize results back to the rest of the population (no pretest) and resulting in decreased external validity. A potential solution for this issue is to use the Solomon Four Group Design⁵, where additional control and experimental groups are added that receive the posttest only.


Positioning of Outcomes Items within CEhp Activities

There are several designs in the collection of data that could be considered for CEhp activities and have a bearing on the sample size. One common design involves placing outcomes question items outside of the educational content, either on printed forms, or if collected via audience response system technology, preceding any informing content (for baseline items) and following all informing content (for post items). A drawback of this approach is that, for live activities, late arrivals and early departures may significantly truncate the potential sample size. There is also risk that the facilitator may not appropriately coach the learners to participate in the outcomes study, failing to adequately address a specific point that is focused on within one of the outcomes items. In addition, question fatigue is a risk in this scenario as the design places a potentially large set of questions in front of the learner at two points in time within the overall activity.

A more subtle design involves framing outcomes items tightly around the informing content, i.e., the content that should impact a learner's answer to the outcomes question item. Using this methodology addresses several issues with the placement of items outside of the content altogether. First, question fatigue is generally decreased. Items are spread out across an activity and are central to the content, making facilitators more apt to speak specifically to the outcomes items. Likewise, learners are more likely to offer a matched response

“The more diverse the sample is, the larger it will need to be to account for the variability and the more confident we are in the result.”

to the items. Late arrivals and early departures are less likely to compromise your sample size as well, since learners are more likely to engage in the “heart” of the content.

For additional illustrated modules on the concepts of sampling data discussed here, check out [Khan Academy](#) (free login required). In the subject box, write “inferential statistics,” and check out the modules on sampling distribution and confidence intervals. 

Forecast of Next Article

In the next article, Gary Bird, PhD, and Sandra Binford, MAEd, will build on the basic points of sampling featured above and focus on the impact of sampling at various time points after an educational intervention.

For Further Reading:

1. Trochim, William M. The Research Methods Knowledge Base, 2nd Edition. <http://www.socialresearchmethods.net/kb/sampstat.php> (version current as of October 20, 2006). Accessed 4/1/15.
2. Lewis-Beck, MS. 2004. The SAGE Encyclopedia of Social Science Research Methods. Sage Publications.
3. Isaac, S. and Michael, WB. 1981. Recommended sample sizes for two different precision levels. Handbook in Research and Evaluation. 2nd Ed. San Diego, EdITS Publishers.
4. Suresh K, Chandrashekhara S. Sample size estimation and power analysis for clinical research studies. Journal of Human Reproductive Sciences. 2012;5(1):7-13.
5. Solomon, RL. 1949. An extension of control group design. Psychol Bull. 46: 137-50.

Impact of Sampling at Multiple Time Points in Measuring Outcomes of Continuing Education in the Health Professions

By Gary Bird, PhD, CME Senior Learning Strategist, American Academy of Family Physicians;
and Sandra Haas Binford, MAEd, Independent Medical Education Designer and Outcomes Researcher

A central goal of continuing education (CE) is to promote measurable, lasting change in the clinical skill of learners. However, outcomes that provide evidence of our educational activity's value heavily depend on the manner and the timing in which they are measured, whether before or after the learning activity. Outcome measurements at multiple time points require careful consideration of not only sample size and composition but also the appropriate positioning of questions to prevent the introduction of bias and to obtain accurate results.

Scope of This Article

The goal of this article is to facilitate an understanding of the relationship between long-term memory and the true outcomes generated by an educational activity. This article will discuss differences in short-term and long-term memory and the benefits of obtaining posttest information at multiple time points following an activity. We will focus on knowledge and competence measurements as defined by Moore et al,¹ because they are fairly easy to measure; strongly aligned with understanding and appropriate application of acquired knowledge by the individual learner; and are not influenced by environmental and patient factors that can complicate the measurement of performance. Additionally, statistical and practical considerations of repeat test sampling will be provided.

Why Repeat Sampling Is Needed

Immediate, post-activity evaluation of knowledge and competence using appropriate tools such as a knowledge assessment test or case vignette only measures short-term memory recall of new information.² It does not indicate extension of knowledge to learned behavior that we seek to change in providing better patient care. Anyone who

has crammed for an exam will have experienced this phenomenon — and its pitfalls. Although cramming may get one through a test that occurs the next day, the trade-off is that much of the information acquired is quickly lost.

Movement toward quality improvement in health care — driven by changes in behavior based on understanding of clinical evidence and demonstration of skill — means that measurement of short-term retention is no longer sufficient for modern CE. However, because of the information loss after an activity,³ it is important to gather and report educational outcomes data at longer time points. Thus, the data we collect should truly reflect a lasting change in learner knowledge or skill and not be simply a short-lived artifact. Follow-up assessment benefits both CE participants (by reinforcing concepts and best practices) and CE providers (by proving pre-educational gaps and illustrating educational effectiveness to stakeholders).

Linking Measurement to Educational Design

The key to cementing learning is providing a robust program of education that reinforces the initial learning installment. A series of linked CE activities offers opportunities for appropriate and accurate measurement. Outcomes data gathered throughout the series informs us of opportunities to tailor further educational events to the evolving needs of learners. Learning occurs only if the content is right and the educational delivery utilized is relevant to the target learner population.

By assessing how learners are doing over time, a CE provider can bring into play additional educational modalities designed to overcome emerging learner needs, while

also reinforcing the concepts already covered. These follow-up experiences lend themselves to the processes that the brain uses to convert short-term memories into long-term, retained information that can be applied to practice, as shown in **Figure 1**. For example, after an initial live event, learners could be invited to attend a small group session at another meeting or to participate in an online activity that engages them in a live, interactive webinar on gaps identified in the first session. Alternatively, the intervention does not have to be certified by an accredited CE provider, and it could simply direct them through an online community-learning forum in which they can focus on the topic and engage in structured learning (for example, problem solving) by interacting with peers and experts.

Follow-up measurement can be either part of an educational initiative (designed to reinforce content and best practices as it gathers educational outcomes data) or completely distinct from the previous activity. Simply sending out a survey via email could generate data that helps describe the evolving needs of the learner. However, when this is done, CE professionals must balance the need to ask enough questions to form valid conclusions with the risk of losing respondents through their lack of motivation to complete a survey.

Content and Use of Questions in Pretesting and Post-Testing

Pretesting at or before the beginning of an activity provides learners with a chance to reflect on their current levels of knowledge, skill and readiness to engage in the activity. Pretest and pre-survey data may also support broader, population-level evidence of knowledge, competence and performance-level gaps documented in the pre-initiative needs assessment.

In considering content for the pretest and posttest questions, there are two choices regarding how to measure key concepts covered in the education. The first, and easiest, option is to repeat a question verbatim in the pretest and posttest, for “repeated measures.” This gives learners, planners and analysts a direct, “apples-to-apples” comparison of the learners’ pre- and post-activity status for change relative to the learning objectives and baseline gaps that need to be addressed and narrowed through education. However, use of the same questions can lead to false positives, as learners may simply remember the answer key from the pretest and repeat it without truly understanding the concepts. A test platform that allows questions to be randomized can help reduce this risk.

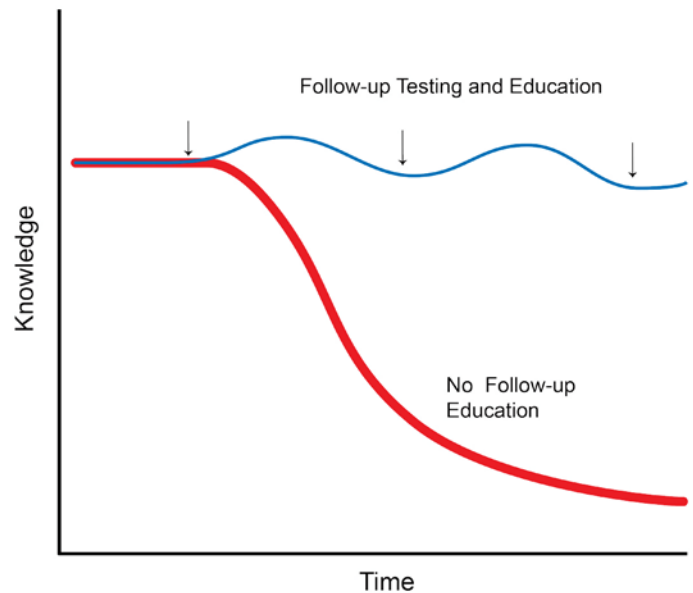


Figure 1. Hypothetical graph of skill decline with time following an initial educational event against the profile that contains multiple testing and linked, individualized education.

Another, and sometimes better, option is to write two questions, one for use in each test, that focus on the same educational point, which should assess a major content item related to a learning objective, as discussed in Article Three (“How to Ask Good Evaluation Questions”) of this series. The question methodology and appearance will be different, encouraging learners to depart from regurgitation and to bridge concepts with experience. The primary caveat with using this methodology is that the CE provider must be careful to craft questions in such a way that the difficulty of the posttest matches that of the pretest. This is not an easy thing to accomplish, as it requires a great deal of skill on the part of test writer. Another concern is the challenge of explaining to reviewers of the outcomes report that using different but equivalent questions was purposeful and constructivist in nature.

Content and Purpose of Follow-up Testing

Conducting additional testing or surveying at later time points (follow-ups) is one of the most important steps for building lasting retention of concepts involved in an educational intervention. Regular testing at multiple time points reinforces learning and motivates the learner to revisit the educational activity and re-engage with concepts. The concepts are now assimilated long term and become a part of their working memory, and therefore are available to improve routine clinical practice. As with the single posttest, the best option is for each test to focus on the concept involved in the learning objective rather than an exact repetition of pretest and posttest items. But again, the same caveats apply.

Table 1. Strengths and Weaknesses of Data Collected from Multiple Posttests

	Paired Data	Unpaired Data (aggregate data)
Strengths of Methodology	<ul style="list-style-type: none"> • Highly robust, lends itself well to statistical analysis • Eliminates effect of outliers 	<ul style="list-style-type: none"> • Easy to obtain • Large sample sizes
Weaknesses of Methodology	<ul style="list-style-type: none"> • Difficult to get the same learners to complete all the tests • Potential of small sample size may prohibit ability to form conclusions from the data • This problem increases as more tests are added 	<ul style="list-style-type: none"> • Outliers can have an impact on the conclusions drawn

Implications of Test Timing for Statistical Analysis

Before the launch of any pretest, posttest or follow-up test instrument, CE professionals need to consider the statistical treatment of the data collected and the sample from which the data were collected. It is important to keep in mind the external validity of the sample as outlined in Article 4 (“Concepts Involved in Sampling Data”) of this series. The sample size may be impacted by repeat measurements, because over an extended period of time, fewer learners will continue to participate in and respond to the testing, bringing up the question of what data you will compare in your analysis. See **Table 1**.

One way to handle shrinking sample sizes is to take all the aggregate data, irrespective of the sample size for each test instrument and compare the results for each test. However, if you take this approach, the samples in these tests are not identical (unpaired samples),⁴ and you must be careful to not let extreme, singular responses — or outliers — alter the interpretation of data from the majority of the sample. This can be dealt with by choosing the correct descriptive statistic to use in your analysis. Alternatively, you may wish to include only individuals who have completed all of the tests in your analysis. By doing this, you will ensure that your sample is the same throughout (paired samples), which makes it less likely that outliers will affect the results. The downside of limiting the dataset to individuals who have completed all instruments is that sample sizes may be too small to allow for robust statistical analysis of the data. Later articles in this series will focus on the statistical methods needed to analyze paired and unpaired samples, as well as how to draw appropriate conclusions from these types of data.

Practical Issues in Follow-up Sampling for Repeated Measures

Including unpaired data in analysis affects both the statistical test that can be used and the validity of the studied data for interpretation. But having zero follow-up data is no longer an option for most CE providers, so conducting repeated measures with the same questions and sample groups is needed. The challenge is to recruit enough responses to follow-up tests that data from early instruments can be analyzed and interpreted alongside frequently fewer responses to later instruments. Recruitment of clinicians who are eligible for inclusion in the study analysis can be costly in terms of human and financial resources, so goals and timing for test and survey recruitment before, during and after the educational activity must be set out and budgeted for in the educational plan before the activity begins.

Yet, having multiple follow-ups may be so challenging in recruitment that data are invalid in one or more test or survey. For example, when an activity has more than one follow-up instrument and fewer responses to the first than the second, portions of the sample in the second instrument would have reinforcement from the first instrument, making respondents to both instruments a different subpopulation. It is often better to put all the necessary follow-up questions into one follow-up instrument so that a snapshot — showing the current status as noted in this series’ Article Three — can be taken and reported with fewer variables and a more informed interpretation.

Case Study

The benefits of multiple posttests can be observed in this example demonstrating the proof of principle for multiple time point testing (retention testing) described above.

The American Academy of Family Physicians (AAFP)⁵ Board Review Self-Study Activity has long used pretesting and post-testing as a means to determine the effectiveness of the education on family physician learner knowledge outcomes. However, until recently, these outcomes were more focused on providing key indicators of immediate learning, rather than expanding the long-term retention of concepts required for learners to cover the evidence-based knowledge scope of family medicine as encountered in their board examination. With this in mind, staff at the AAFP began to fundamentally alter the structure of the activity to consider linking testing with a personalized education approach.

The AAFP Board Review Self-Study Activity is an online educational program in which learners can access on-demand sessions originally recorded from a live board review activity. The cardiology data — an aggregate of sessions for five different topics — is shown in Figure 2. Each session in the cardiology topic group utilized five identical pretest and posttest questions, based on the content in each presentation for a total of 25 questions. Staff expanded the testing in the retention test so that one month prior to the start of the American Board of Family Medicine Board spring examination period in April 2015, all learners who had purchased the activity were invited to retake the test. For learners to be included in the data gathering for this test, they had to wait at least three weeks between taking the posttest and retention tests. Although retention test questions were identical to the pretest and posttests, the order in which they were asked was randomized in an attempt to limit bias. Once learners had completed the retention test, they were directed back into the activity and invited to re-engage in cardiology sessions where they had shown a persistent knowledge deficit.

A paired approach to the analysis of data was utilized to eliminate the effect of outliers on test score averages — thus, using this methodology, learners had to have taken all three tests to be counted in the sample. A total of 50 learners completed the cardiology pretest, posttest and retention tests. This represented approximately 6 percent of the learners, which demonstrates how difficult it can be to obtain robust sample sizes in this type of measurement. **Figure 2** shows the aggregate scores for learners in each of the three tests. Pre-testing indicated learners had a moderate level of baseline knowledge in cardiology; a deeper analysis

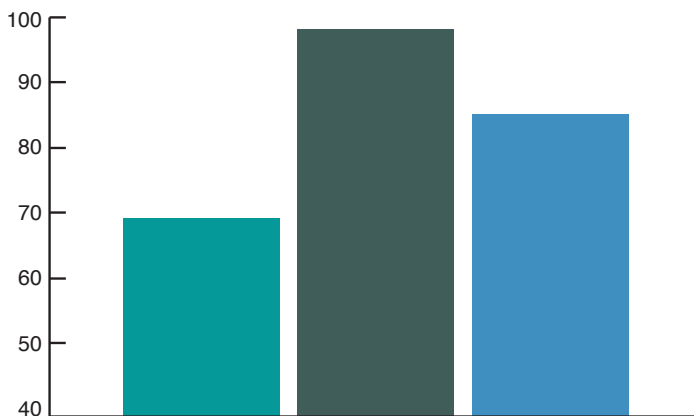


Figure 2. Average pre-, post- and retention test score for physician learners engaged in cardiology topics in the AAFP Board Review Self-Study Activity

of the individual sessions (data not shown) revealed specific topic strengths and weaknesses. As expected, immediate post-testing showed a high score based on immediate regurgitation of knowledge with a drop in the average score in the retention test. The retention scores suggest learners have retained a large proportion of the information covered in the content in readiness for their exam, although without additional educational interventions, this score would be expected to further decline over time. [*](#)

Forecast of Next Article

In the next article, Melanie Bird, PhD, and Derek Dietze, MA, FACEHP, CCMEP, will provide foundational concepts for understanding different types of data sets and levels of measurement and an introduction to descriptive statistics used to analyze that data.

References

1. Moore DE Jr, Green JS, Gallis HA. Achieving Desired Results and Improved Outcomes: Integrating Planning and Assessment Throughout Learning Activities. *J Contin Educ Health Prof.* 2009 Winter;29(1):1-15.
2. Burkiewicz JS, Bruce SP, Weberski JA, Ritter JL, Sohn AH. Pre- and Post-Rotation Assessment of Pharmacy Student Learning. *J Pharm Teaching.* 2005;12(2):83-96.
3. Cowan N. What Are the Differences Between Long-term, Short-term, and Working Memory? *Prog Brain Res.* 2008;169:323-338.
4. Definitions of Paired and Unpaired data samples. www.tufts.edu/~gdallal/paired.htm
5. American Academy of Family Physicians Home Page. www.aafp.org/home.html

Basic Concepts of Data Sets

By Melanie D. Bird, Ph.D, American Academy of Family Physicians; Derek T. Dietze, MA, FACEHP, CHCP, Improve CME, LLC

Data collection and interpretation is essential to understanding outcomes and change due to educational activities. Appropriate outcome measurement is dependent on the correct amount and type of data being collected. This article lays the foundation for understanding the different types of data and the terminology and methods to begin to analyze that data.

Scope of This Article

The goal of this article is to provide an overview of types of data collected from CEhp activities and the types of simple descriptive analyses that can be performed. This article will discuss qualitative and quantitative data, levels of measurement and simple descriptive statistics. We will also pay attention to differences in paired and aggregate data as well as a discussion of scale measurement.

What Do We Mean by 'Data?'

Dictionaries define data as facts or figures from which conclusions can be drawn.¹ Data are categorized in a variety of ways starting with a designation of qualitative or quantitative.

- Qualitative data are nonnumeric (words, not numbers).¹ Qualitative data is also termed categorical as it is descriptive in nature falling into distinct categories such as color, texture, appearance or in terms of CEhp participant demographics (sex, type of practice, healthcare specialty). Responses to open-ended questions collected from evaluation forms and surveys, and answers to questions administered to focus groups and in phone interviews are also considered qualitative data. Qualitative data can be described but not measured until linked to a numerical scale, becoming quantitative.
- Quantitative data is numerical data that can be measured. These data can be discrete with finite values (whole numbers) or continuous with infinite possibilities resulting in decimals (1.9, 2.0, 2.1, 2.2, 2.3, etc).¹ Examples of questions that collect qualitative and quantitative data are shown in **Table 1**.

Table 1. Examples of Questions that Collect Qualitative and Quantitative Data

Questions that Collect Qualitative (nonnumerical) Data	Questions that Collect Quantitative (numeric) Data
What is your profession? A. Physician B. Physician Assistant C. Nurse Practitioner D. Other (please specify): _____	Rate your level of agreement with the following statement: “[insert statement]” 1=Strongly disagree 2=Disagree 3=Neutral 4=Agree 5=Strongly Agree
What is your primary specialty? a. Cardiology b. Endocrinology c. Internal Medicine d. Other (please specify): _____	How many patients with asthma do you encounter each week? # of patients: _____
[Multiple choice knowledge question] a. Distractor 1 b. Correct Answer c. Distractor 2 d. Distractor 3	How confident are you in your ability to identify patients with testosterone deficiency? 1=Not confident at all 2=Not very confident 3=Somewhat confident 4=Very confident 5=Extremely confident
What do you plan to change in your practice based on your participation in this CME activity?	How often do you now plan to use the “squeeze test” when you suspect rheumatoid arthritis? 1=Never 2=Not Often 3=Sometimes 4=Often 5=Always
Based on any changes you have made in your practice since completing the CME activity 8 weeks ago, what outcomes have you observed in your patients?	For how many patients with COPD do you believe you have provided improve care since completion of the CME activity 6 weeks ago? # of patients: _____

Another way to describe data collected in CEhp activities is whether the data are paired or aggregate.

- Paired data occur when each point in a data set is matched to a data point in a second data set.² This occurs routinely with pretest and posttest data. For instance, a participant’s response to each pre question is matched to their response to the same post question, or a participant’s score on the pretest is matched to his or her score on the posttest. For each question, you will have the same amount of pre as post respondents.
- Aggregate data are not matched, and you will likely have a different number of participants answering each pre question and the same post question. Examples of paired and aggregate data from a live activity are shown in **Figures 1 and 2**.

It is important to note that regardless of the type of data you are collecting, you are still working with a sample. (See articles number four and five of this series on sampling). When the data from the entire sample or population are described, the number of participants is designated by N. For a subset of the sample or population, the number of data points or participants is designated with n. For example, for the paired data shown in **Figure 1**, N=10 (the total number of paired responses), and in **Figure 2**, there are N=8 pre responses and N=5 post responses. If you only wanted to analyze the physicians who answered questions in **Figure 1**, n=6 because it is a subset of the sample.

Types of Scale Data

When discussing qualitative and quantitative data, there are four levels of measurement: nominal, ordinal, interval and ratio.³⁻⁴ The complexity and types of analysis increase as we progress from one level to the next.

- Nominal data are the first level of measurement. Nominal data are qualitative data that are contained in mutually exclusive categories. In essence we are classifying the data. This can be done using letters, words or even numbers depending on the classification. For example, we can classify participant gender using words — male or female or using letters — M or F. Other examples of nominal data might be type of practitioner, specialty, education level or even geographic location.
- Ordinal data found in the second level of measurement are very similar to nominal data except that there is an ordered relationship between the numbers or items. However, the interval between units is not meaningful. An example seen in CEhp activities might be a ranking of interest level for a potential activity (first, second or third) or the level of participant satisfaction.

Keypad#	Profession	Specialty	Practice Setting	PtsSeenPerWk	PreQ1	PreQ2	PreQ3	PreQ4	PostQ1	PostQ2	PostQ3	PostQ4
23	Physician	FM	Group	20-30	a	c	b	3	a	b	a	4
99	Physician	IM	Private	10-20	a	a	c	4	a	a	c	5
43	NP	FM	Group	20-30	d	a	d	3	a	b	c	3
26	PA	FM	Clinic	1-5	c	a	a	3	b	b	c	4
75	Physician	IM	Private	>40	b	b	a	5	a	b	c	5
67	Nurse	ICU	Hospital	1-5	b	b	c	4	a	b	c	5
89	Physician	OB/GYN	Private	10-20	a	c	a	3	b	b	c	3
44	NP	FM	Group	1-10	a	d	a	3	a	a	c	4
14	Physician	IM	Hospital	10-20	a	a	a	5	a	c	b	4
55	Physician	IM	Clinic	30-40	c	a	b	4	a	b	c	4

The same participants take a test at two discrete times, “Pre” and “Post”.

Figure 1. The same participants take a test at two discrete times, “pre” and “post.”

Pre1OrPost2	Profession	Specialty	Practice Setting	PtsSeenPerWk	Q1	Q2	Q3	Q4
1	Physician	IM	Private	20-30	a	b	c	3
1	NP	FM	Group	10-20	a	b	d	4
1	PA	FM	Group	20-30	b	a	d	3
1	Nurse	ICU	Hospital	1-5	c	d	c	4
1	Physician	OB/GYN	Private	>40	d	b	a	5
1	Physician	FM	Private	1-5	a	b	c	3
1	Physician	FM	Private	10-20	a	b	c	3
1	PA	IM	Group	1-10	a	a	a	5
2	Physician	FM	Group	10-20	a	b	c	4
2	Physician	FM	Group	30-40	a	a	c	5
2	NP	FM	Group	20-30	b	b	d	5
2	Physician	IM	Private	30-40	a	b	c	5
2	Physician	FM	Clinic	>40	a	b	c	4

Two groups of participants (“Pre1” and “Post2”) take the same test.

Figure 2. Two groups of participants (“Pre1” and “Post2”) take the same test.

- Interval data are the third level of measurement. These are data that are classified with a specific order and a defined interval or spacing between units. The distances or intervals are equivalent; however, there is no zero point. Interval measurement in the CEhp world might include the level of confidence a participant has in his or her ability to complete a specific patient care task from 1 to 5. The difference in confidence of participants at 2 and 3 is the same as participants with scores of 4 and 5.
- Ratio data are similar to interval data but have a clear definition of zero such as height or weight. These data can also be viewed against each other as a ratio. A score on a pretest of nine is three times higher than a score of three.

Simple Descriptive Statistics for Data Sets

Descriptive statistics are used to describe the data without drawing any conclusions beyond the simple parameters or characteristics of the data.⁵ Inferential statistics are used to expand on the immediate data and attempt to generalize to a bigger population. Later articles in this series will explain inferential statistics in more detail and how use them successfully to demonstrate success in educational outcomes for CEhp activities. For descriptive statistics, the type of parameter used will vary based on the type of data. See **Table 2** on page 9.

Nominal data are categorical and can be described using frequencies (counts) or percentages. For example, a CEhp activity was conducted with 90 participants who were then asked to provide demographic data including age, sex and specialty. The answers would then be counted and the frequency or percentage can be reported as follows in **Table 3**. Frequency or percentage is also used to describe ordinal data. For these data, CEhp providers would show the number of participants rating a CEhp activity or rating their level of satisfaction with the activity. An outcome measure for an activity may be said to achieve greater than 90 percent satisfaction.

Ordinal data can also be described or displayed as frequency or percentage. Participants may be asked to rank several CEhp topics in order of importance. Researchers can then use this information to inform programming. For example, if 75 percent of participants rank hypertension as being more important than cholesterol to them, then CEhp providers might consider creating more activities focused on hypertension. Interval and ratio data can be analyzed in more detail with more advanced descriptive and even inferential statistics.

Interval and ratio data exist along a scale with established spacing between the values and a distribution is created of data points across the scale. The distribution of the data set is a listing of all the possible values or intervals and how often they occur in the sample (population). In other words, it is the spread of frequencies or percentages. In our example above, the frequency of participants who fall into each category can be shown graphically with a bar graph (histogram), with bars for each category whose height represents the number of participants in that category (**Figure 3**). Often, CEhp providers are concerned with the central tendency value, which is the middle (typical) value in the data set. It can be measured using the mean, median or mode. All three parameters can be observed visually in a graph of the distribution.⁶ The distribution can be spread evenly, skewed left or right or be mixed. When the mean, median and mode of a data set are all equal, the result is a normal distribution. If one of the parameters is found within a higher value (median and mode greater than mean), the distribution will be skewed left (negatively skewed), for a lower value (mean greater than median and mode), it will be skewed right (positively skewed). There are instances even when the data can have two most popular values, resulting in a bimodal distribution (i.e., a bump on the left and on the right).

Measures of Central Tendency

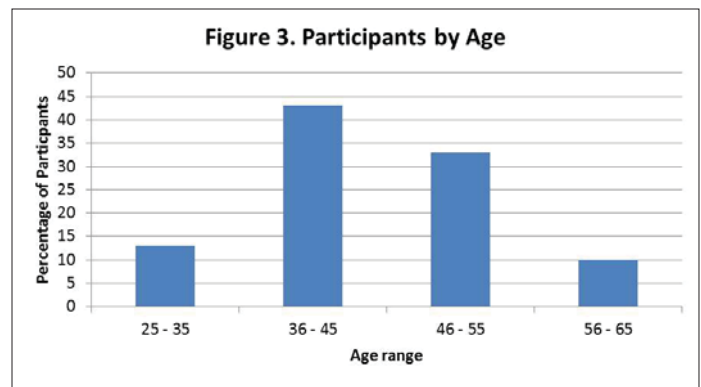
As outlined above, each of the levels of measurement can be described with different statistical values based on the

Table 2. Types of Data and Descriptive Statistics Used in CEhp Activities

Type of Data	Definition	Example	Descriptive Statistics Used
Nominal	Categorical	Participant sex, age, specialty	Mode
Ordinal	Ordered (ranked) or directional	Level of satisfaction for event	Median
Interval	Equal intervals with no zero	Participate level of confidence in knowledge of subject matter	Symmetrical distribution: Mean Skewed distribution: Median
Ratio	Equal intervals with zero	Participant pretest and posttest scores (paired or aggregate)	Symmetrical distribution: Mean Skewed distribution: Median

Table 3. Descriptive Statistics for Nominal Data

	Data	Frequency	Percentage (%)
Age Range	25-35	12	13
	36-45	39	43
	46-55	30	33
	56-65	9	10
Sex	Male	40	44
	Female	50	56
Specialty	Family Medicine	35	39
	Pediatrics	27	30
	Internal Medicine	21	23
	Geriatrics	7	8




type of data and distribution. Each statistic focuses on a particular value in the data set, as shown on **Table 4** on page 10. Certainly most readers will be familiar with these terms, but every introduction to statistics includes a review of mean, median and mode. The most common measurement of central tendency is the mean or average value of your data set. It is calculated by dividing the

sum total by the number of data points. You may see the symbol μ used for the mean of the population x or for the mean of the sample. It is important to remember that the mean is only valid for interval and ratio data and can be influenced by extreme data points (outliers). Therefore, it is important to consider any outliers in your data set and if they should be included or if they are the result of bias in the sample or experimental design.

The median is the middle value of the data after it is sorted into ascending order. There would be an equal number of data points above and below the median. For example, in the set (2, 5, 8, 9, 10, 22, 23), the median = “9.” If the number of data points is an even number, the median is defined as the mean of the middle two values. The median can be used for interval, ratio and ordinal data and in contrast to the mean, it is not influenced by outliers. Therefore, the median may be more appropriate to use when the mean may be distorted with extreme data points. For instance, this is commonly the case for home values and income data.

The mode is the value that is most popular or occurs most frequently in the data set. If the data is graphed to show the frequency distribution, the mode would be the peak. The mode can be used with any of the data types described above. The mode is used a lot with nominal data as it will show the most popular answer. However, the mode may be less useful for interval and ratio data if the data distribution is spread thin resulting in no data points having the same value.

Forecast of Next Article

In the next article, Tanya Horsley, PhD, and Gary Bird, PhD, will expand on the fundamental concepts addressed in this article and discuss in greater detail the distribution and variation of data sets. 

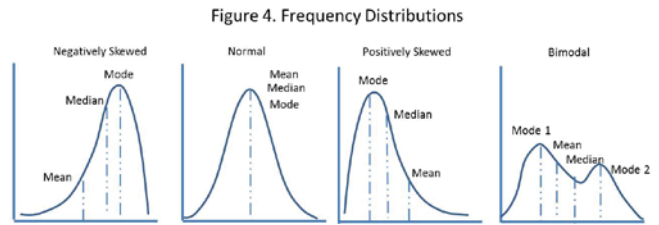


Table 4. Definitions of Mean, Median and Mode

Descriptive Statistic	Definition
Mean	Average value of data set
Median	Middle value of data set when ordered from low to high
Mode	Most popular value (most frequent) in the data set

References

1. Calkins, K.G. 2005. Definitions, uses, data types, and level of measurements. *Applied Statistics*. <http://www.andrews.edu/~calkins/math/edrm611/edrm01.htm>. Accessed June 2015.
2. http://stattrek.com/statistics/dictionary.aspx?definition=paired_data.
3. Trochim, W.M. 2006. Levels of Measurement. *The Research Methods Knowledge Base*. <http://www.socialresearchmethods.net/kb/measlevl.php>. Accessed June 2015.
4. <http://www.graphpad.com/guides/prism/6/statistics>. Accessed June 2015.
5. Trochim, W.M. 2006. Descriptive Statistics. *The Research Methods Knowledge Base*. <http://www.socialresearchmethods.net/kb/statdesc.php>. Accessed June 2015.
6. <https://www.mathsisfun.com/data/standard-normal-distribution.html>. Accessed June 2015.

Distribution and Variation in Data Sets

By Gary C. Bird, PhD, American Academy of Family Physicians; Melanie Bird, PhD, American Academy of Family Physicians; and Sandra Haas Binford, MAEd, Full Circle Clinical Education, Inc.

Introduction

As outcomes data obtained in assessing CE activities are analyzed, it quickly becomes clear that there is typically not one value or answer that may be associated with all learners, but that data are often spread across a range of possible values. In statistics, data for a complete population are rarely available, so a smaller “sample” is often used to proxy for the entire population. However, in order to make inferences about data for an entire population of learners, the properties of the data sample must be understood. The description of a data set provides foundational information that may allow researchers and their readers to generalize findings for the sample to the entire population of learners.

Scope of Article

The goal of this article is to provide an overview of the statistical terms encountered in describing a sample of data for one variable. There are two large groups of statistical data — “descriptive” and “inferential”— and this article discusses the descriptive statistics that must be summarized about the samples before researchers and readers may make any valid inferences about findings. Readers should be able to use terms describing a data set; doing so is essential to explaining the value of the research to larger populations than those learners who were sampled in a study.

Three Fundamental, Descriptive Parameters of Data Sets

The description of a sample data set includes parameters, which show the location, shape and reach for all data points for a single variable. When data in a set are limited to one variable (i.e., the answers to one question), we call the set “univariate.” In an assessment of univariate data sets, the parameters are typically the following:

1. The distribution
 - a. Bell curve or normal distribution
 - b. Skew

2. The central tendency
 - a. Mean
 - b. Mode
 - c. Median
3. The dispersion
 - a. Range
 - b. Variance
 - c. Standard deviation
 - d. Standard error
 - e. Confidence interval/interquartile range

These terms are used in everyday language and often have very different meanings than those specified for statistical purposes, so definitions and examples are provided below. Another frequently heard term is “data points,” which are referred to as “values” below.

Physician Age	Frequency (%)
<36	10
36-45	21
46-55	42
56-65	19
>65	8

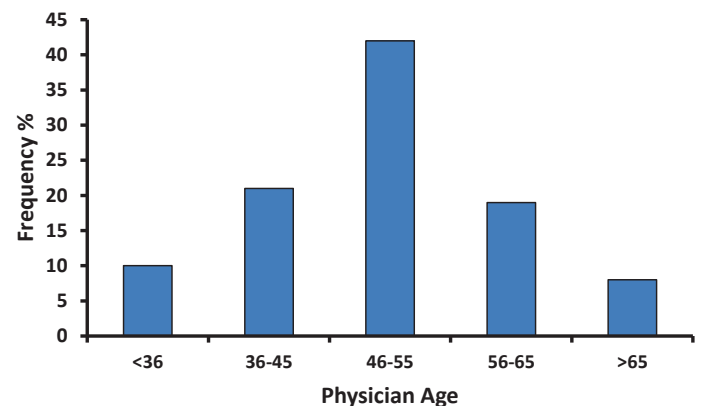


Figure 1. shows examples of frequency distribution depictions from an educational event in which the grouped age ranges of physician learners are shown against percentage frequency.

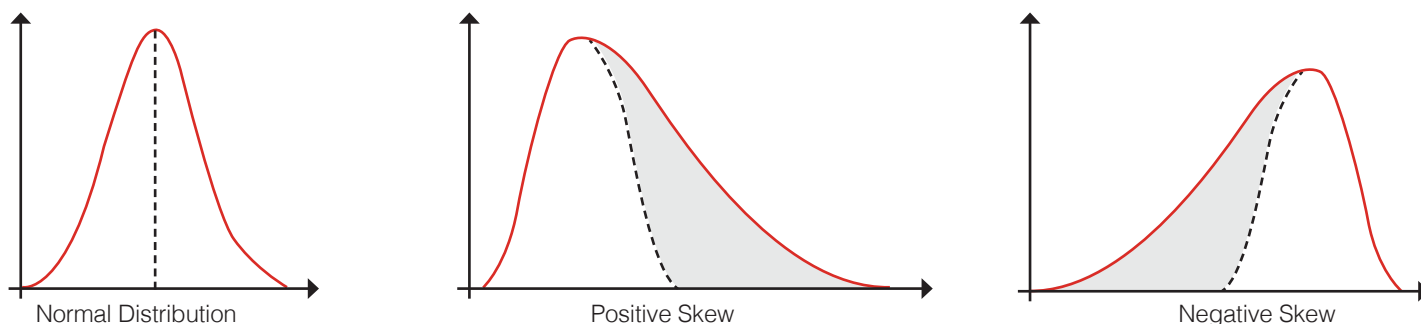


Figure 2. Distribution of Data.

Distribution of the Data Set

As discussed in the previous article, “Basic Concepts of Data Sets,” distribution is a summary of the frequency of individual values (data points) or ranges of values for a variable.¹ One of the most common ways to describe a single variable is with a frequency distribution. Depending on the variable measured, all of the data values may be represented or they may be grouped into categories first (e.g., as value ranges). The frequency of each variable can be calculated as either a simple count or a percentage value. Frequency distributions can be depicted as a table or a graph (Figure 1).

Although the table in Figure 1 is useful in detailing specific data, its more meaningful distribution profile becomes truly apparent when the data are graphed. In graphs, the data take on a distinctive shape based on the frequency values of the group. The shape or profile of the distribution can be matched with an idealized form with a very large sample size, and typically one of three profile types will appear. The first is called the “normal” (or standard) distribution (Figure 2), in which the peak of the distribution visually falls at the midpoint of available values with two entirely symmetrical “tails” on either side. Readers may have heard of a “bell curve,” which is another term for the normal distribution.² Sometimes data are obtained in which the majority of the sample leans away from the midpoint. When the majority of data fall toward the left side of the graph’s x (horizontal) axis, and a longer tail is seen on the right, a second profile appears called a “right (or positive) skew” distribution (Figure 2). Conversely, when a majority of data fall on the right of the x axis and a long tail occurs on the left, a third profile is observed, called a “left (or negative) skew” distribution (Figure 2).

Central Tendency and Distribution of Data

The concepts of the central tendency¹ — with terms defining the mean, median and the mode of a data set — play an important role in descriptive statistics. We reviewed these in

Example 1. Calculation of the mean, median and mode (3 Measures of Central Tendency)

The following post-test data set was obtained from a group of eight learners (N = 8):

12, 19, 12, 14, 18, 15, 12, 22

Mean

$$(12 + 19 + 12 + 14 + 18 + 15 + 12 + 22) \div 8 = 124/8 = \mathbf{15.5}$$

Median

12, 12, 12, **14, 15**, 18, 19, 22 = **14.5** (midpoint of 14 and 15)

Mode

12, 12, 12, 14, 15, 18, 19, 22 = **12**

the previous article, and we will return to them many times in this series. As described in “Basic Concepts of Data Sets,” the mean or average value of the data set is calculated by dividing the sum total by the number of data points. The median is the middle value of the data after it is sorted into ascending order, while the mode is the value that is most popular or common, occurring most frequently in the data set.

When the numerical values for central tendency are matched against the three types of distribution curve noted above, a consistent pattern emerges that helps define the data profile. In a normal distribution with no skew, the mean, mode and the median are always equal. In right (positive) skewed distributions, the mean has the largest value, followed by the median, and the mode will have the smallest value. Conversely, in left (negative) skewed distributions, the mode has the largest value, followed by the median, while the mean has the smallest value. Even when values are not plotted on a graph, seeing these patterns can tell the researcher or reader that data are skewed and the sample may not represent the larger population.

It is tempting to look at the data in Example 1 and conclude it has characteristics that appear to match a right skewed profile. Although this is true, it should also be noted that often marginal differences in mean, median and mode that are present in

small sample sizes disappear when the sample size is increased. This is because the impact of “outlier” or aberrant single data points decrease in large populations, and therefore generally, the normal distribution is the one that is most commonly observed.

Dispersion of the Data Set

Dispersion¹ refers to the spread of the values around the central tendency (mean, median or mode). There are two common measures of dispersion: the range and the standard deviation. The range is simply the highest value minus the lowest value. From the data given in Example 1, the range would be $22 - 12 = 10$. Range can be misleading in interpreting the relevance of a data set to populations that are larger than the sample, because range can be increased greatly if just one person in the sample has a much higher or much lower value than the others, called an “outlier.”

The standard deviation¹ shows the relationship between the mean of the data set and all of the points in a data set. It is a more accurate and detailed estimate of dispersion because, unlike the range, outliers do not have an exaggerated impact on the value obtained. The standard deviation also efficiently summarizes dispersion in graphs, tables and text-based results, offering layout advantages for reporting without distracting the reader’s eye from the core data and without omitting essential data needed for later meta-analyses or any assessment of a study’s quality.

In addition to expressing the variability of a population, the standard deviation is commonly used to measure confidence in statistical conclusions. For example, the margin of error in polling data is determined by calculating the expected standard deviation in the results if the same poll were to be conducted multiple times. The reported margin of error is typically about twice the standard deviation. Calculation of the standard deviation is most simply done using a computer program. Microsoft Excel, for example, will quickly calculate the standard deviation for a data set. However, the calculation of the standard deviation is not difficult, and it is worth examining how this value is derived (**Example 2**). A lay article on why the standard deviation is important is [located online](#).

Using an online standard deviation calculator may help you get more familiar with understanding standard deviation and its relation to the mean.

Standard Error

As discussed in Article 4, “[Concepts Involved in Sampling Data](#),” the spread of scores across an infinite number of samples would be the standard error (sampling error or SE).¹ The SE is related to standard deviation and the size of your sample and is

Example 2. Calculation of the Standard Deviation

Using the post-test data from Example 1, which were 12, 19, 12, 14, 18, 15, 12, 22, and had a mean of 15.5

- I. First, calculate the deviation of each post-test score from the mean, and square the result of each. These are the results you would get:

$$15.5 - 12 = 3.5; 3.5^2 = 12.25$$

$$15.5 - 19 = -3.5; -3.5^2 = 12.25$$

$$15.5 - 12 = 3.5; 3.5^2 = 12.25$$

$$15.5 - 14 = 1.5; 1.5^2 = 2.25$$

$$15.5 - 18 = 2.5; 2.5^2 = 6.25$$

$$15.5 - 15 = .5; .5^2 = 0.25$$

$$15.5 - 12 = 3.5; 3.5^2 = 12.25$$

$$15.5 - 22 = -6.5; -6.5^2 = 42.25$$

- II. Next, calculate the mean of the squared values.

$$12.25 + 12.25 + 12.25 + 2.25 + 6.25 + 0.25 + 12.25$$

$$+ 42.25 / 8 = \mathbf{12.56}$$

(note this value is also known as the sample variance)

- III. Finally, take the square root of the variance to give the sample standard deviation.

$$\sqrt{12.56} = \mathbf{3.5}$$

Example 3. Calculation of Standard Error

$$3.5 / \sqrt{8} =$$

$$3.5 / 2.8 = 1.25$$

calculated by dividing the standard deviation by the square root of the sample size (n). In Example 3 below, we calculate the SE for this sample of posttest results.

The lower the standard deviation and the larger the sample size, the smaller the sample error and the more reliable the result becomes. Therefore, the SE is a measurement of reliability.

For example, if we increased the sample size to 15, and the mean had a standard deviation of 3.1, the SE would equal $3.1 / \sqrt{15} = 3.1 / 3.9 = .8$. We would be able to state that this sample is a more reliable estimate of this particular variable in the larger population than the previous sample with the SE of 1.25.

Confidence Intervals

The standard error can then be used to construct confidence intervals. Confidence intervals give a range of values where we can assume the true value will be found. Most of the time, 95 percent is used to set the intervals. In order to determine the 95 percent confidence intervals (CI), two standard errors are added and subtracted from the mean. Using the data from Example 1, the 95 percent CI for

those values would be from 12.8 to 18.2. This means that for our sample, we can be 95 percent confident that the true mean lies somewhere between 12.8 and 18.2.

Percentiles and Quartiles

Performance is a key focus for CE professionals, as it is an indicator of how successfully an educational activity prepared the learner. Learners also like to see percentiles reported, as it shows how well they performed compared to their peers. Percentiles and percentile ranks are used to provide performance indicators.⁴ These are terms that relate one learner's performance to the larger group. A common example of the use of percentiles can be seen in national standardized college entrance exams and medical specialty board examinations. Percentiles are a means of dividing a distribution into two or more groups based on the rank desired. It is important to note the distinction between percentile and percent. Being in the 90th percentile does not mean that the learner correctly answered 90 percent of the questions. Instead it means the learner scored better than 90 percent of her fellow participants.³

Percentiles are calculated by ordering the values in a data set from smallest to largest and then multiplying the total number of values by a particular percent. This is the number or position of the value that represents the proverbial line in the sand. Continuing with the data from above, Example 4 shows how to determine the 80th percentile for those scores.

The use of the term “quartile” is often used to describe the 25th, 50th and 75th percentiles. The 25th percentile is also known as quartile 1, or the lower quartile, the 50th percentile is known as quartile 2 and is equal to the median, while the 75th percentile is known as quartile 3, or the upper quartile. The interquartile range, also sometimes called the “middle fifty” is the first quartile subtracted from the third quartile.

Normality and Use of Parametric and Non-Parametric Tests

Up to this point, we have focused on descriptive statistics used to describe the simple parameters or characteristics of the data. The next step in analysis is to use inferential statistics to expand on the immediate data and attempt to generalize to a bigger population. The distribution of the data (described above) determines which statistical test will be most appropriate.⁴ For data that follow a normal distribution, parametric tests are used. A parametric test is a test that assumes a normal distribution across the population and that the measures are from an equal interval scale. These tests are reserved for specific data types such as interval and ratio data. Examples of parametric tests include t-tests and analysis of variance (ANOVA), which will be discussed later in this series. For other data types like

Example 4. Calculation of Percentiles

To calculate the 80th percentile

- I. First order the scores: 12, 12, 12, 14, 15, 18, 19, 22
- II. Next multiply the number of scores ($n = 8$) by .8 (80%), which is 6.4. Round this number up to the next whole number (7).
- III. Then count the values from left to right to find the 7th score, which is 19.
- IV. The score of 19 represents the 80th percentile.

nominal or ordinal data that may not be distributed normally, nonparametric tests are used. Nonparametric tests are those used to analyze data that have a skewed distribution or when the outcome has limits of detection or outliers. Examples of nonparametric tests include chi-square, the Mann-Whitney test and the Fisher Exact Test.

Conclusion

Knowing descriptions of data and seeing a normal distribution of variable data means that one can confidently generalize findings for the study's sample as being relevant to larger groups of the studied, target population. If means and standard deviations of data, which can be pictured in a curve, do not match the data that one expects of a very large learner population, then the findings of the study should not be generalized. A study's significance tests (p values), confidence ranges and effect sizes matter little if the integrity of the sample is not assured from careful demographic filtering and indications that data lie in a normal distribution. 📊

Forecast of Next Article

In the next set of articles, Derek Dietze, MA, FACEHP, CHCP, and Erik Brady, PhD, CHCP, will offer insight in how to analyze pre- and post-activity data. These articles will discuss inferential statistics and their use in different scenarios of CEhp activities.

References

1. Trochim, W.M. 2006. Descriptive Statistics. *The Research Methods Knowledge Base*. Available at <http://www.socialresearchmethods.net/kb/statdesc.php>
2. <https://en.wikipedia.org/wiki/Skewness>
3. Rumsey, D.J. What percentile tells you about a statistical value. *Statistics for Dummies*, 2nd Ed. Available at <http://www.dummies.com/how-to/content/what-percentile-tells-you-about-a-statistical-value.html>
4. Boston University of Public Health. 2013. Nonparametric Tests. http://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/BS704_Nonparametric/BS704_Nonparametric2.html



Beginner's Guide to Measuring Educational Outcomes in CEhp

How to Analyze Your Baseline, Post-Activity Change Data

Part 1: Baseline, Post-Activity Multiple-Choice Questions

By Erik D. Brady, PhD, CHCP, Wake Forest Baptist Medical Center; and Derek T. Dietze, MA, FACEhp, CHCP, Improve CME, LLC

This article addresses ACEhp National Learning Competency:

- [Competency Area 3.i: Measuring the Performance of Activities and the Overall Program](#). Use evaluation and outcomes data ... (C) Analyzing assessment data in order to draw conclusions about the effectiveness of the activity/intervention based on expected results.

One of the most common types of outcomes data that CEhp professionals can work with comes from multiple-choice knowledge and competence questions asked both before and after a CEhp activity. These data are typically collected at the time of the activity via paper forms, online or an Audience Response System (ARS). Summarizing the results for each baseline to post-activity question and calculating a “*P* value” for the change in number of correct answers from baseline to post-activity can provide insights into the effectiveness of your CEhp activity. It can also enhance the credibility of your outcomes reports, and provide a foundation for improving future activities.

This article focuses on providing a working definition of *P* value and provides step-by-step directions on how to calculate a *P* value for baseline to post-activity multiple-choice knowledge and competence questions. Two cases are highlighted: the first addresses collected non-paired data, and the second highlights collected paired data (for more information about paired and non-paired data, see the article “[Basic Concepts of Data Sets](#),” published in the September 2015 issue of the *Almanac*).

What is a “*P* value”?

A *P* value (the “*P*” means “probability”) is generated from a test of statistical significance (a mathematical formula).¹ In the case of comparing baseline answers to posttest answers of multiple-choice questions, the *P* value indicates

whether or not the before-to-after change in correct answers was statistically significant. Simply put, the *P* value represents the role that chance plays in your outcomes.

The calculation used for *P* value results in a value between 0 and 1 and can be interpreted.² In general, a *P* value of .05 or less represents the “gold standard” in scientific research, meaning that 95 percent of the time your findings are statistically significant. This means that there is only a 5 percent likelihood that a calculated change from baseline to post-activity would occur by chance alone if the same education were offered to additional learners of similar demographics.

Statistical significance does not necessarily mean practical significance. Only by considering context can you determine whether a difference is practically significant (that is, whether it requires action).¹

In general, if there is an increase in correct answers from baseline to post-activity, you want to see a *P* value of 0.05 or lower in order to state in your outcomes report, “There was a statistically significant increase in correct answers from baseline to post.”

- A small *P* value (typically ≤ 0.05) indicates strong evidence that the baseline to post change is real and is not due to chance. An *increase* in correct answers from baseline to post with a *P* value of ≤ 0.05 is a *positive* result — a statistically significant increase in correct answers. A *decrease* in correct answers from baseline to post, with a *P* value of ≤ 0.05 is a *negative* result — a statistically significant decrease in correct answers.
- A large *P* value (> 0.05) indicates weak evidence that the baseline to post change is real, and it is more likely due to chance. An increase in correct answers from baseline to post, with a *P* value > 0.05 means that while more people answered correctly post than at

baseline, the increase was not statistically significant. Conversely, if there was a decrease in correct answers with a *P* value of >0.05, that decrease was not statistically significant or meaningful.

Sometimes analysts will refer to a “null hypothesis” and an “alternative hypothesis”² when conducting tests of statistical significance. In the context of baseline/post-activity multiple-choice questions, the null hypothesis is that *there is no difference* between correct answers baseline and post-activity. The statistical test determines if this null hypothesis is correct or not. If you get a *P* value of <0.05, then you reject the null hypothesis and accept the alternative hypothesis, which is that *there is a difference* between correct answers baseline and post.

Case: Unpaired Baseline/Post Multiple-Choice Question Data

In this case study, assume that from your hospital grand rounds CME activity you collected participants’ answers to six multiple-choice baseline questions before the activity and the same questions post-activity. You have a stack of completed baseline questionnaires and a stack of post questionnaires, and there are no names on the questionnaires so you cannot match them. Also, you have 38 completed baseline questionnaires and 31 completed post questionnaires because some participants left early and did not complete the post questionnaire. How do you determine if there was a statistically significant increase in correct answers for each multiple-choice question?

Step 1: Enter your data into Excel.

Table 1 shows what your data should look like in Excel after initial data entry. Due to space limitations in this article, we are only showing results from the first eight completed baseline questionnaires and the first five completed post questionnaires. Also, we show only data for three of the six questions. Notice that beside each column where you have entered each participants’ answer to a question (a, b, c, or d), you have “coded” their answer as either correct (1) or incorrect (0). Since the correct answer for question 1 is B, you have labeled the column “Q1CorrectB” to help with your coding.

Step 2: Summarize the number of correct and incorrect responses in a table.

The remainder of these steps focuses on question one results. You would repeat these steps for each of the six questions. After doing your data entry for all 38 baseline questionnaires and all 31 post questionnaires for baseline question 1, you count up the 25 participants who answered correctly and the 13 who answered incorrectly.

Table 1. Unpaired Data Entry and Correct Answer Coding

Pre or Post	Q1	Q1 Correct B	Q2	Q2 Correct A	Q3	Q3 Correct D
Pre	a	0	a	1	d	1
Pre	b	1	d	0	d	1
Pre	b	1	c	0	c	0
Pre	b	1	a	1	b	0
Pre	c	0	a	1	c	0
Pre	b	1	d	0	d	1
Pre	d	0	d	0	d	1
Pre	c	0	b	0	d	1
Post	b	1	a	1	d	1
Post	c	0	d	0	d	1
Post	b	1	a	1	c	0
Post	b	1	a	1	d	1
Post	b	1	c	0	d	1

Table 2. Question 1 Baseline/Post Correct/Incorrect Answers

	Correct	Incorrect
Pre	25	13
Post	27	4

Table 3. Blank 2x2 Contingency Table

	Outcome 1	Outcome 2
Group 1		
Group 2		

For post question one, 27 answered correctly and four incorrectly. Using this information, in Excel, create **Table 2**. Notice that you have used the count of correct/incorrect answers, not percentages.

Step 3: Enter results in an online tool to calculate the P value.

Proceed to a free online statistics tool to enter your data. While many are available, **GraphPad** is a simple one to use. **Table 3** shows a simple table (called a “2x2 contingency table”) as shown on the Web page where you will enter your data. Type in and replace “Outcome 1” with “Correct,” “Outcome 2” with “Incorrect,” “Group 1” with “Pre” and “Group 2” with “Post.” Then enter the data from the Excel table you created in Step 2.

What you entered should now look like **Figure 1**, and as shown, you select “Chi-square without Yates’ correction” as the test you want completed, select “Two-tailed” and press “Calculate.” If any numeric value you enter into the table (as shown in Figure 1) is five or less, it is recommended that you select “Fisher’s Exact Test” under “Which Test,” instead of the Chi-square test.

Step 4: Review results and create a significance statement.

The *P* value calculated using this method is 0.041, as shown in **Figure 2** highlighted in yellow. Thus, your statement regarding question one would be, “There was a statistically significant increase in correct answers from baseline to post ($P=0.041$, baseline $n=38$, post $n=31$, Chi-square test).”

Finally, showing percent correct baseline and post in a figure summarizing question results is recommended. For example, for question one, 65.8 percent (25/38) answered correctly at baseline, and 87.1 percent answered correctly at post (27/31). Thus, the absolute increase from baseline to post was 21.3 percent (87.1 percent minus 65.8 percent). However, it is more common to state the relative increase, which would be 32.4 percent, using the formula: $[(87.1-65.8)/65.8] \times 100$. An online calculator for this can be found at Marshu.com.

Working with Paired Data

Having a data set in which the responses to multiple-choice items are assigned to specific individuals is definitely a preferred situation. Such a scenario allows you to consider data from only those learners that offered a response to a question at baseline and at post-activity. Working with a set of data that is restricted in this way, is called working with “paired data.” Generally, statisticians think of this as cleaner data that allows for a more powerful analysis to definitively quantify change.

As with unpaired data, the first step is to calculate the group baseline correct percentage and the group post-activity correct percentage to determine the delta for the group being considered. At that point, however, a distinct test is required to calculate the *P* value. As was shown with unpaired data, the best way to describe the calculations is to show an example.

Case: Paired Baseline/Post Multiple-Choice Question Data

In this case, assume a recent data set for your online educational activity had five outcomes questions that were asked within the delivery of content to assess changes in competence. Learners were able to respond to question items as they desired, but the data analysis was restricted to only those who offered a response to both a baseline and a post question for

Figure 1. Completed Table and Selection of Test and Tails

	Correct	Incorrect
Pre	25	13
Post	27	4

Which test

There are three ways to compute a *P* value from a contingency table. Fisher’s test is the best choice, as it always gives the exact *P* value, while the Chi-square test only calculates an approximate *P* value. Only choose Chi-square if someone requires you to. The Yates’ continuity correction is designed to make the Chi-square approximation better. With large sample sizes, the Yates’ correction makes little difference. With small sample sizes, Chi-square is not accurate, with or without the correction.

- Fisher’s exact test (recommended)
- Chi-square with Yates’ correction
- Chi-square without Yates’ correction

A *P* value can be calculated with either one or two tails. We suggest always using two-tailed (also called two-sided) *P* values.

- Two-tailed (recommended)
- One-Tailed

Calculate

Figure 2. *P* Value from Chi-square Test

Analyze a 2x2 Contingency Table

	Correct	Incorrect	Total
Pre	25	13	38
Post	27	4	31
Total	52	17	69

Chi-square without Yates correction

Chi-square equals 4.174 with 1 degree of freedom.

The two-tailed *P* value equals 0.0410

The association between rows (groups) and columns (outcomes) is considered to be statistically significant.

each question item. The resulting data was found across the activity, as shown in **Table 4** (see page 9).

All changes appear positive, and you shared them with the course director. The course director then indicates a desire to understand the statistical significance of these findings.

Step 1: Access a statistical computation tool.

In order to determine a *P* value for paired data, several tests are available. An easy one to use with free access is found at [Graph-Pad](#). In order to access the appropriate tests to analyze the data found in Table 4, go directly to the [McNemar’s test Web page](#)

on GraphPad. **Figure 3** shows the screen that will appear to assist in your calculation of a *P* value using paired data.

Step 2: Organize your data.

In order to put your data into this tool, a bit of data organization is required. There are four possible results when a learner responds to a multiple-choice question twice. First, the learner can answer incorrectly (I) at baseline and correctly (C) at post; for use of this tool, this highly desirable outcome is referred to as “Control = No and Case = Yes.” The first cell in the GraphPad tool is for the number of times that this situation occurred. Second, the learner can answer correctly (C) at baseline and incorrectly (I) at post; the number of times that occurs goes in the second cell of the tool corresponding to “Control = Yes and Case = No.” Third, the learner can answer correctly (C) at baseline and correctly (C) at post (“Control = Yes and Case = Yes”). The number of times this “reinforcement” finding occurs goes in the third cell in the tool. Finally, a learner can answer incorrectly (I) at baseline and incorrectly (I) at post (“Control = No and Case = No”), and the number of times that occurs goes in the fourth and final field in the tool. A click on “Calculate” returns the *P* value for paired data, as well as several other pieces of information.

To see how this functions, **Table 5** shows the four different scenarios described above for the five questions presented in **Table 4**. “No/Yes” refers to the count of individual learners who missed the question at baseline but selected correctly at post.

Step 3: Load your data and execute calculation.

It may take a bit of time to prepare your data for the calculation, but once a table like **Table 5** is created, plugging the data into GraphPad is fairly simple. An example is shown for Question 1 in **Figure 4** (see page 10).

The key value is the “two-tailed *P* value” determined as 0.6069 for the example Question 1, which is shown framed in a red box in **Figure 4**. When McNemar’s test is performed for all five example questions, we can add *P* values to our original table, shown as in **Table 6** (see page 10).

It is necessary to verify that the *number of discordant pairs is greater than 20* in order for this calculation to be valid. That value can be found in the summary narrative from the GraphPad calculation tool, shown framed in a yellow box in **Figure 4**. This is an important distinction, as Question 1 has only 34 discordant pairs, even though the response count (*n*) is 123.

Step 4: Analyze your change.

While your first glance at the data showed positive change on all items, when you consider the *P* values, you find that the significance of the calculated change from baseline to post

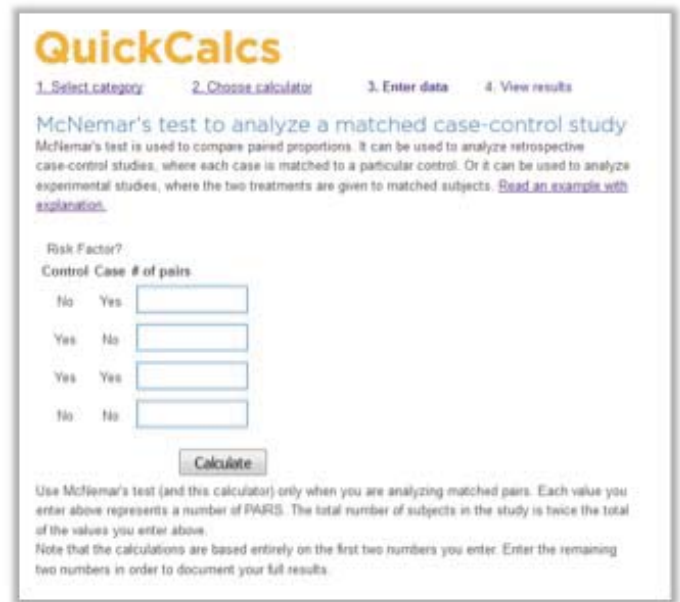
Table 4. A set of paired data from a typical educational activity; *n* = number of learners responding to both the baseline and post instance of the question

Question #	n	Baseline average correct	Post average correct	Change (D)
1	123	76%	80%	+4%
2	119	51%	77%	+26%
3	51	51%	88%	+37%
4	36	17%	33%	+16%
5	36	53%	94%	+41%

Table 5. Data from **Table 4** organized to show the count of learners according to the four different possible ways that multiple-choice items can be answered twice

Question #	n	No/Yes	Yes/No	Yes/Yes	No/No
1	123	19	15	79	10
2	119	38	7	54	20
3	51	21	2	24	4
4	36	7	1	5	23
5	36	15	0	19	2

Figure 3. McNemar’s Test Input Screen



(deltas) are greatly varied. For example, for Question 1, you see 123 total matched responses — an “*n*” that you might expect to give rise to a significant finding. While the delta is only 4 percent, you might be tempted to say that the change is significant and would be greater if there was a lower baseline. However, the *P* value does not confirm that analysis. Conventional criteria would suggest that this positive 4 percent change is fairly random and meaningless.

Question 2, on the other hand shows a highly statistically significant finding. Large deltas, when combined with large n's, typically do lead to a statistically relevant finding.

Questions 3, 4 and 5 are also included here for specific reasons. With Question 3, the number of discordant pairs is 23 (n=51), barely allowing for validity of the calculation. What that indicates is that many learners didn't change the way that they responded to this question item, so it's fair to say that many of the 51 percent of learners who got the question correct at baseline had their choice reinforced.

Question 4 has only 8 discordant pairs (n=36) and a P value that falls slightly higher than the threshold (< 0.05, as previously mentioned) that most use to qualify for statistical significance. If the activity is ongoing, it may be wise to await additional data. This type of finding is sometimes referred to as a "change trending toward significance."

Question 5 looks very significant with a +41 percent% delta and a P value of 0.0003. Unfortunately, the number of discordant pairs is 15 (n=36), which falls short of the needed 20. Because of the P value, you might describe this as a "change that is likely to reach significance with additional sampling." As with Question 4, waiting for additional data may address this, if the possibility of additional data collection exists.

Limitations

For unpaired data sets, there are definite limitations when the n of the baseline and post groups are highly varied. In that case, it's possible that the two groups may not be an accurate reflection of each other. For example, consider a scenario where the baseline group has 150 responses and the post group has 30 responses. In addition, the 30 post responses are all members of your target audience, but the 150 baseline responses are a mix of target audience and non-target audience. It's possible that the calculated delta and P values may be less valid than your calculations would suggest. This is one of the rationales for using paired data whenever possible.

For paired data, we mentioned several times that McNemar's test has a minimum number of discordant pairs limitation. There are other calculations that avoid this specific limitation, but for the purposes of this beginner's data analysis article, we've chosen to propose the use of McNemar's test as it covers most cases and generally works quite well when paired n's are higher than at least 40 on an individual question.

Summary

The P value represents the role that chance plays in your outcomes. Researchers accept that chance may play some role

Figure 4. A sample calculation in GraphPad with the results page shown at right

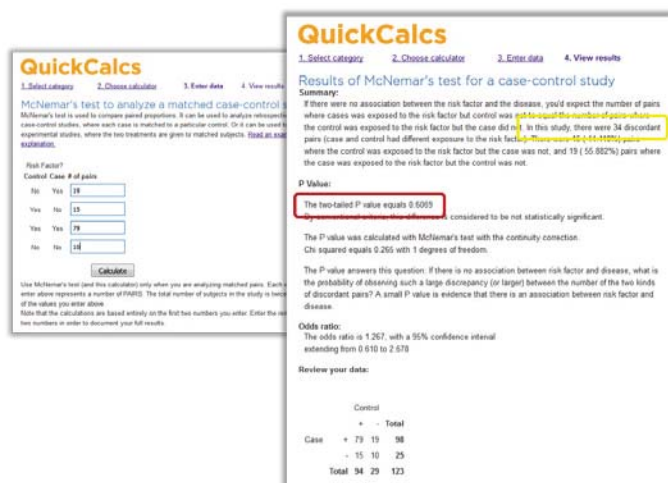


Table 6. Addition of calculated P values to assess significance of change percentages (Δ)

Question #	n	Baseline average correct	Post average correct	Change (Δ)	P Value
1	123	76%	80%	+4%	.6069
2	119	51%	77%	+26%	<.0001
3	51	51%	88%	+37%	.0002
4	36	17%	33%	+16%	.0771
5	36	53%	94%	+41%	.0003

in their findings, but only if that chance is 5/100 (P = 0.05) or less. Any greater likelihood of something happening due to chance is grounds for saying that your findings are random.

With the right tools, setting up and calculating the statistical significance of your findings is really fairly simple, and you can make it easily repeatable if you have the inclination to better understand what your data have to say. One topic that hasn't been addressed elsewhere in this article is the issue of very low n's; what if you have very few respondents (e.g., 15 at baseline and 10 at post)? Low participation in analysis of multiple-choice outcomes data does present a challenge that is not easy to overcome.

In short, the lower the number of your respondents, the larger the change from baseline to post in correct answers needed to achieve statistical significance. In the absence of enough data (typically no fewer than 30 responses at baseline and/or post is needed to give yourself a chance at measuring significance using a credible statistical test like Chi-square), our recommendation is to show relative increase in correct answers from baseline to post. 🌱

References

1. <http://www.dummies.com/how-to/content/statistical-significance-and-pvalues.html> Accessed 10/23/15.
2. <http://www.dummies.com/how-to/content/what-a-pvalue-tells-you-about-statistical-data.html> Accessed 10/23/15.

Resources

1. Jason Olivieri, CMEPalooza, Statistical Analysis in CME Outcomes, <http://cmepalooza.com/march21/statistical-analysis-in-cme-outcomes-olivieri/>
2. Erik D. Brady, PhD, CHCP, CMEPalooza "Excel"lent Tricks for the Non-Expert: Exploring the Beauty of the Cells. <https://www.youtube.com/watch?v=11I75UrlqxE>

*Your digital marketing partner
in a modern healthcare world*

mms provides email and marketing solutions
that effectively deliver your message to
healthcare professionals.

AMA Physicians · AAPA Physician Assistants
Nurse Practitioners · Pharmacists · Hospital Managers
Email · Direct Mail Marketing

 **mms**
message delivered.
mmslists.com • 800.MED.LIST

Beginner's Guide to Measuring Educational Outcomes in CEhp

How to Analyze Your Baseline/Post-Activity Change Data

Part 2: Baseline/Post Rating Scale (Ordinal) Questions

By Erik D. Brady, PhD, CHCP, EDBPHD Consulting, and Derek T. Dietze, MA, FACEHP, CHCP, Improve CME, LLC

In our last article, we addressed how to analyze results from multiple choice questions asked both before and after a CEhp activity. Another common type of baseline/post-activity question involves the use of a rating scale to assess changes in confidence, agreement or frequency of use. Again, these data are collected at the time of the activity on paper forms or through an Audience Response System (ARS). Summarizing the results for each baseline/post-activity question and calculating a “*P* value” for the change in ratings can provide insights into the effectiveness of your CEhp activity. It can enhance the credibility of your outcomes reports and provide a foundation for improving future activities. Finally, with appropriate goal statements within your mission statements, these types of measures can be a powerful way to analyze your overall CEhp program.

Scope of this Article

This article focuses on providing step-by-step directions on how to calculate a *P* value for baseline/post-activity rating scale questions. The results data from these questions are considered “ordinal” data—they have an order (i.e., lowest to highest). We have highlighted two cases: The first addresses a situation in which you have collected paired data, and the second addresses a case in which you have collected unpaired data (see “Basic Concepts of Data Sets” in the [September 2015](#) issue of the *Almanac*). You will also find it helpful to review the concepts of data that are “normally” or “not normally distributed,” how to choose a “parametric” or “non-parametric” statistical test (see “Distribution and Variation in Data Sets” in the [October 2015](#) issue of the *Almanac*), and the definition of a “*P* value” (see “How to Analyze Your Baseline/Post Activity Change Data Part 1: Baseline/Post Multiple Choice Questions in the December 2015 *Almanac*”).

Working with Paired Data

As with multiple choice items, having a data set in which the

Table 1. A set of paired data from a typical educational activity; *n* = number of learners responding to both the baseline and post instance of the question

Question #	<i>n</i>	Baseline Average (Current Use)	Post Average (Planned Use)	Change (Δ)
1	21	2.57	5.33	+2.76
2	48	5.63	5.98	+0.35
3	56	5.43	5.70	+0.27
4	32	3.47	5.91	+2.44
5	17	5.59	5.65	+0.06

responses to rating scale items are assigned to specific individuals is definitely a preferred situation, as it allows us to work with a paired data set. As with unpaired data, the first step is to calculate the group baseline average and the group post average to determine the “delta,” or “change,” for the group being considered. Note that the symbol for delta is Δ . At that point, we run a distinct test to calculate the *P* value. As we have done for multiple choice items (in the previous article) and with unpaired ordinal data (later in this article), we show the best way to describe the calculation using an example.

Case 1: Paired Baseline/Post-Activity Rating Scale Question Data (Parametric)

A data set for your online educational activity that was recently conducted had five rating scale questions that were asked after the delivery of content to assess changes in intent (competence) for specific practice strategies that were supported by the activity content. A seven-point semantic differential scale was used (descriptive words only at each end of the scale, but not for the values two through six) in each case: 1 = No Use and 7 = Extensive Use. Learners were able to respond to question items as they desired, but the data analysis was restricted to only those who offered a response to both a baseline and a post question for each question item. The resulting data was found across the activity, as shown in [Table 1](#). All changes

appear positive, and when you share them with the course director, he/she indicates a desire to understand the significance of these findings.

Step 1: Access a Statistical Computation Tool

In order to determine a *P* value for paired ordinal data, several tests are available. One challenge of working with ordinal data is that you need to understand whether or not your data are parametric (i.e., shaped like a Bell curve) or non-parametric (i.e., not shaped like a Bell curve). A t-test, which we describe below, is a parametric test.

When we graph our ordinal data, particularly when we ask learners to rate something (e.g., confidence or intent-to-use practice strategy), it is not atypical to find that our data resembles a Bell curve, or would resemble a Bell-shaped curve if the data were extrapolated. If that is the case, then using a paired t-test is totally appropriate.¹ If the data looks flat or in some way does not resemble a Bell curve, then we are in a position where we need to use a non-parametric test. In the case of paired ordinal data, the Wilcoxon signed-rank test is the most appropriate test to use.¹ We will direct readers to easy online tools for both the t-test and the Wilcoxon test, and you can use a free online tool from [Social Science Statistics](#).

An easy tool for the paired t-test can be found at [GraphPad](#). As with all the tools that we refer to, access is free and available online. In order to access the appropriate tests to analyze the data found in Table 1, access the website, select “Continuous Data” on the first screen, click “Continue,” then select “t-test to compare two means” and click “Continue,” or go directly to the [Web page](#). **Figure 1** shows the screen that will appear to assist in your calculation of a *P* value using paired ordinal data.

Step 2: Organize Your Data and Execute Calculation

In general, it’s typically easiest to select the second radio button, “Enter or paste up to 2000 rows,” unless you have a very small data set and prefer to manually key in the data. You will need to ensure that each row in your data set represents an individual learner. “Group One” is the baseline cohort, so all baseline data should be copied and pasted into the first column shown in Figure 1. “Group Two” is the post cohort. The tool does allow you to replace the labels with “Baseline” and “Post” if you so choose. Once you add your data to the Group One and Group Two columns, select “Paired T-test” under “3. Choose a Test.” Click “Calculate Now” under “4. View the Results” to return the *P* value for your paired data set, as well as several other pieces of information. An example of the returned data set is shown in **Figure 2** for Question 1.

Figure 1. T-test input screen on GraphPad

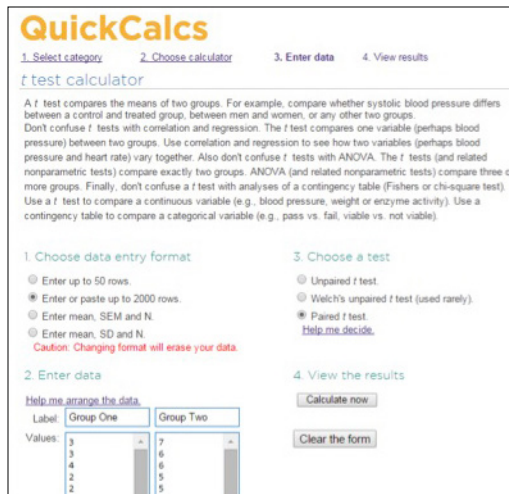


Figure 2. A sample calculation in GraphPad with the results page shown



Table 2. Addition of calculated P values to assess significance of change percentages (Δ).

Question #	n	Baseline Average (Current Use)	Post Average (Planned Use)	Change (Δ)	P value
1	21	2.57	5.33	+2.76	<.0001
2	48	5.63	5.98	+0.35	.0022
3	56	5.43	5.70	+0.27	.0799
4	32	3.47	5.91	+2.44	<.0001
5	17	5.59	5.65	+0.06	.5795

The key value is the “two-tailed *P* value,” determined as < 0.0001 for the example Question 1, which is shown framed in a red box in Figure 2. When the t-test is performed for all five example questions, we can add *P* values to our original table, shown as in **Table 2**.

Step 3: Analyze Your Change

While your first glance at the data showed positive change

on all items, when you consider the *P* values, you find that the significance of the calculated change from baseline to post (deltas) are greatly varied. For example, for Question 1, you see only 21 total matched responses – an “n” that you might expect to push the boundaries of a significance calculation. Clearly, this is a highly significant finding, even with a fairly low number of responses. Question 4 is similar; baseline and post means are higher and are across a larger number of respondents than Question 1, and we find a similar *P* value.

Question 2 has a larger number of responses, but the change from baseline to post is smaller than what was found for Question 1. As was noted in the last article in the series, this *P* value (0.0022) is still lower than the threshold used by most to qualify for significance at the 95 percent confidence level (< 0.05).

Questions 3 and 5 are also included here for specific reasons. With Question 3, the number of responses is 56, which is the largest n in the data set, and yet, we find a *P* value of 0.0799, which is considered not statistically significant, but may be described as trending toward significance. A higher number of respondents may lead to a *P* value < 0.05. Question 5 has only 17 responses, barely allowing for validity of the calculation. The *P* value for Question 5 is 0.5795, also an insignificant finding. The core message is that simply increasing the number of respondents doesn’t always lead to a significant finding, but this must be considered on a case-by-case basis.

Case 2: Paired Baseline/Post-Activity Rating Scale Question Data (Non-parametric)

If shortly after completing your analysis from the previous case you take a look at a bar chart representation of your data and find that it does not appear to be parametric, you may question the *P* values that you calculated using the t-test. In order to feel greater confidence, you decide to re-analyze your original data set using a non-parametric test. The Wilcoxon signed-rank test is an appropriate test for paired ordinal data that are not normally distributed.¹

Step 1: Access a Statistical Computation Tool

An easy tool for the Wilcoxon signed-rank test can be found on the [Social Science Statistics website](#). **Figure 3** shows the screen that will appear to assist in your calculation of a *P* value using paired data.

Step 2: Organize Your Data and Execute Calculation

As with the tool available on GraphPad, the Wilcoxon signed-rank tool requires data to be formatted into baseline (Treat-

Figure 3. Wilcoxon signed-rank test input screen on Social Science Statistics

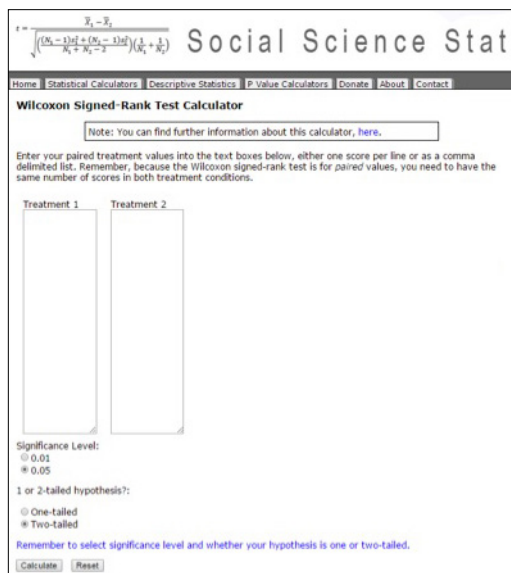
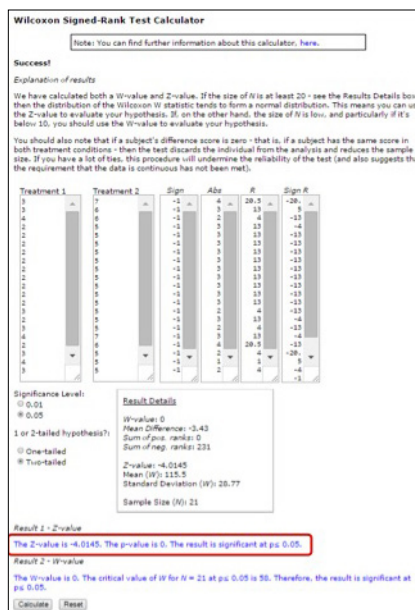


Figure 4. A sample Wilcoxon calculation with the results page shown



ment 1) and post (Treatment 2) groups. Data are pasted or hand-keyed into each box. Once “Significance Level: 0.05” and “Two-tailed” hypothesis have been indicated, click “Calculate.”

The key value is the “Z-value,” determined as “0” for the example Question 1, which is shown framed in a red box in **Figure 4**. It is not atypical to show “0” as “<0 .0001” to show a consistent precision in the analysis of *P* values. When the Wilcoxon signed-rank test is performed for all five sample questions, we can add *P* values to our original table, as shown in **Table 3**.

Step 3: Analyze Your Change

In general, the interpretation of your results is the same regardless of whether you select a parametric or a non-parametric test of significance for ordinal data. Comparing the calculated *P* values reveals minimal changes for findings that are highly significant. For example, for Questions 1 and 4, both parametric and non-parametric tests result in a *P* value <0.0001 regardless of the chosen test. Question 2 also falls into the category of significant with either test.

Question 3, however, presents an interesting finding, in that the parametric test suggested near significance (*P* = 0.0799), and the non-parametric test (*P* = 0.0173) falls into the range of significance. The non-parametric test would allow you to call this significant, because we found that our data was not Bell shaped in this case.

Question 5 has only 17 responses, barely allowing for validity of the calculation with the t-test (*P* = 0.5795). The Wilcoxon signed-rank test cannot be performed for samples that are this small, having a requirement of 17 non-zero differences in order to be valid.

Case 3: Unpaired Baseline/Post-activity Rating Scale Question Data (Parametric)

From your hospital grand rounds CME activity, you collected participants' answers to four confidence rating scale baseline (pre) questions before the activity and the same questions post-activity (i.e., please rate your confidence in your ability to XYZ: 1 = Not confident at all, 2 = Not very confident, 3 = Somewhat confident, 4 = Very confident, 5 = Extremely confident). You have a stack of completed baseline questionnaires and a stack of post questionnaires, and there are no names or email addresses (unique identifiers) on the questionnaires, so you cannot match them. Also, you have 38 completed baseline questionnaires and 31 completed post questionnaires, because some participants left early and did not complete the post questionnaire. How do you determine if there was a statistically significant increase in ratings for each confidence question?

Step 1: Enter Your Data into Excel

Table 4 shows what your data should look like in Excel after initial data entry. Due to space limitations in this article, we are only showing results from the first eight completed baseline questionnaires and the first five completed post questionnaires.

All changes appear positive, and when you share them with the course director, he/she indicates a desire to understand the significance of these findings.

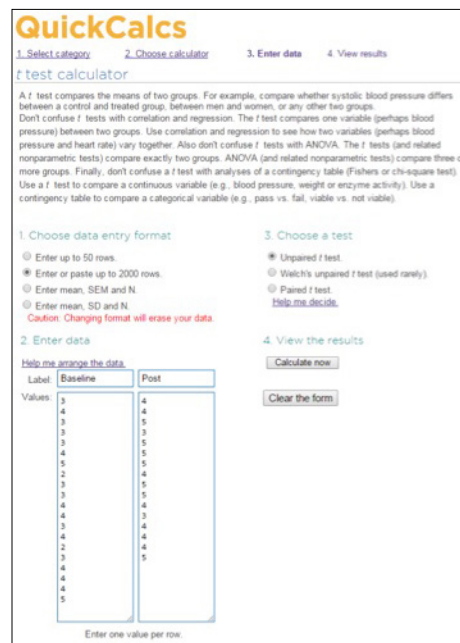
Table 3. Addition of calculated P values to assess significance of change percentages (Δ)

Question #	n	Baseline Average (Current Use)	Post Average (Planned Use)	Change (Δ)	P value
1	21	2.57	5.33	+2.76	<.0001
2	48	5.63	5.98	+0.35	.0050
3	56	5.43	5.70	+0.27	.0173
4	32	3.47	5.91	+2.44	<.0001
5	17	5.59	5.65	+0.06	N/A

Table 4. Unpaired Data Entry in Excel

Baseline or Post	Q1	Q2	Q3	Q4
Baseline	3	1	4	3
Baseline	4	2	5	3
Baseline	4	3	5	3
Baseline	3	2	5	5
Baseline	3	3	5	3
Baseline	5	4	4	3
Baseline	4	3	4	3
Baseline	3	3	3	4
Post	3	2	5	4
Post	4	2	5	5
Post	4	3	4	5
Post	5	3	5	4
Post	5	4	5	5

Figure 5. Unpaired t-test data entry screen in GraphPad



Initially you make the assumption that your data are normally distributed (Bell-shaped). The appropriate test for normally distributed unpaired ordinal data is the unpaired t-test, a parametric test.¹

Step 2: Access a Statistical Computation Tool and Enter Your Data

As with the paired data analysis, the t-test is still an appropriate test for unpaired data, so you used the same tool found at [GraphPad](#). Under “1. Choose Data Entry Format,” select “Enter or Paste up to 2000 Rows,” and under “3. Choose a Test” select “Unpaired T-test” as shown in **Figure 5**. Label the first column “Baseline” and the second column “Post,” then for confidence question one (Q1) in your Excel file, copy and paste the raw baseline data into the first column, and the raw post data into the second column. For space reasons, not all data are shown in the figure, but for Q1 in our example case, you have 38 data points in the first column and 31 in the second.

Step 3: Calculate the P Value for Q1, Repeat for Q2-Q4 Under “4. View the Results,” click “Calculate Now.” **Figure 6** shows the results with a two-tailed P value of 0.0036 (see the red box), indicating that there was a statistically significant increase in confidence ratings for Question 1 from baseline (3.50/5) to post (4.31/5), $P = 0.0036$, baseline $n = 20$, post $n = 16$, unpaired t-test. Repeat the same procedure to obtain P values for questions two through four, and summarize your results in a table, much like that shown for the paired data cases above.

Case 4: Unpaired Baseline/Post Rating Scale Question Data (Non-parametric)

Shortly after completing your analysis, you find that your data does not appear to be parametric, which makes you question the P values that you determined using the unpaired t-test. In order to feel greater confidence, you decide to re-analyze your original data set using a non-parametric test. The most appropriate test for unpaired ordinal data that are not normally distributed is the Mann-Whitney U test.¹

Step 1: Access a Statistical Computation Tool and Enter Your Data

An easy tool for the Mann-Whitney U test can be found at [Social Science Statistics](#). **Figure 7** shows the screen that will appear to assist in your calculation of a P value using unpaired ordinal data. The Mann-Whitney U test tool requires data to be formatted into baseline (Population 1) and post (Population 2) groups. Data are pasted or hand-keyed into each box.

Step 2: Calculate the P Value for Q1, Repeat for Q2-Q4 With “Significance Level: 0.05” and “Two-tailed” hypo-

Figure 6. A sample calculation for the unpaired t-test in GraphPad with the results page shown

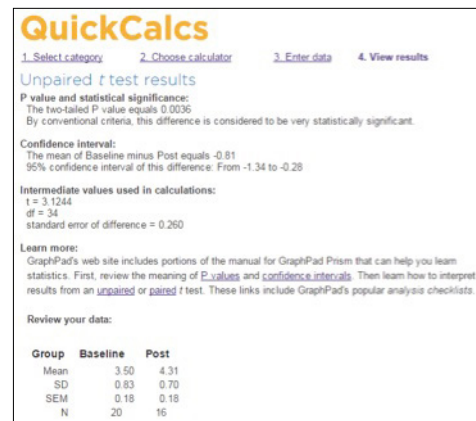


Figure 7. Mann-Whitney U-test input screen on Social Science Statistics



Figure 8. A sample Mann-Whitney U test calculation with the results page shown




esis indicated, click “Calculate.” The results are shown in **Figure 8**. A P value of 0.00854 is shown in the red box, indicating that there was a statistically significant increase in confidence ratings for question one from baseline (3.50/5) to post (4.31/5), $P = 0.0085$, baseline $n = 20$, post $n = 16$, Mann-Whitney U test. Repeat the same procedure to obtain P values for questions two through four, and summarize your results in a table. Review your results as described in the paired data case. An alternative online tool for the Mann-Whitney U test can be found at [Vassar Stats](#) by clicking on “Ordinal Data” on the menu at the left of the page and then on “Mann-Whitney Test.”

Limitations

Some of the same limitations that we described in the previous article on multiple choice items continue to hold true with ordinal data, and we restate them here. For unpaired data sets, there are definite limitations when the n of the baseline and post groups are highly varied. In that case, it’s possible that the two groups may not be an accurate reflection of each other. For example, consider a scenario where the baseline group has 150 responses and the post group has 30 responses. In addition, the 30 post responses are all members of your target audience, but the 150 baseline response are a mix of target audience and non-target audience.

It’s possible that the calculated delta and P values may be less valid than your calculations would suggest. This is one of the rationales for using paired data whenever possible.

Open-access online statistical test tools can be used to calculate P values for your paired or unpaired ordinal data (i.e., rating scale data) that you collect to assess the effectiveness of your CEhp activities. For paired ordinal data, the paired t -test is best when the data are normally distributed, and the Wilcoxon signed-ranks test is best when the data are not normally distributed. For unpaired ordinal data, the unpaired t -test is best for normally distributed data, and the Mann-Whitney U test is best when the data are not normally distributed. 

Reference

1. How to Choose a Statistical Test. <http://www.graphpad.com/support/faqid/1790/> Accessed 12/12/15.

Resources

2. Jason Olivieri, CMEPalooza, Statistical Analysis in CME Outcomes, <http://cmepalooza.com/march21/statistical-analysis-in-cme-outcomes-olivieri/>
3. Erik D. Brady, PhD, CHCP, CMEPalooza “Excel”lent Tricks for the Non-Expert: Exploring the Beauty of the Cells. <https://www.youtube.com/watch?v=11I75UrlqxE>



REACH
HEALTHCARE PROFESSIONALS
FAR & WIDE

We get your message delivered.
You get responses.

CONTACT US AT 800.MED.LIST OR SALES@MMSLISTS.COM

 **mms**
message delivered.
mmslists.com • 800.MED.LIST

Beginner's Guide to Measuring Educational Outcomes in CEhp

Understanding the Impact of Data and Analysis at the Population Level

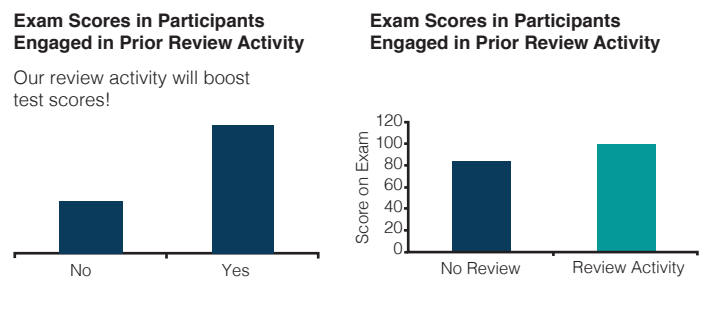
How Common Statistical Mistakes Impact Data Interpretation

By Gary C. Bird, PhD, American Academy of Family Physicians; and Melanie D. Bird, PhD, American Academy of Family Physicians

One of the first things to consider as you review data obtained from a CE activity is to ask the question, “Is the data reproducible?” meaning that another person can construct the same study and obtain the same results. This is a primary concern for randomly controlled trials (RCTs) seeking to tightly control factors that may influence results and, in doing so, allow for robust data collection and statistical analysis that yield definitive answers to a defined research question. Reproducibility is also important for generalizing data across a population. If an RCT is reproducible (and the sample is random), the results should hold for a larger population. Unfortunately for us as providers of CE, the data we obtain from our educational activities can never involve such tight control. CE learners have different motivations for engaging in an activity and different levels of engagement during the activity, resulting in potential sources of variability in the data obtained. Our learners, even when converged by profession or specialty, often vary dramatically across a spectrum in these variables.

This is not to say that data from CE activities has no place in a truly scientific assessment of the impact on CE outcomes. There is considerable knowledge to be gained from collection and analysis of activity data. The education we provide is an important element in the continuous professional development of medical professionals and what we do has the potential to positively impact the way they practice medicine, which ultimately results in the increased health of their patients. Therefore, drawing conclusions from data obtained as a result of our education that carries over to larger populations is the key to not only improving the quality of education, but also improving the means by which we disseminate ideas to others in our area.

Figure 1. Examples of good and bad graphs.



Scope of this Article

Previous articles in the statistics series have talked about specific statistical methodologies that are pertinent to those trying to make sense of their data. In this article, we will go through some of the common mistakes and misconceptions when individuals use these methodologies to analyze, view and interpret CE activity data. We will also highlight key areas that can cause issues with drawing conclusions on a population level.

Inappropriately Graphing or Charting Your Data

For most people, numbers alone generally do not “pop” or liberate trends in the data to provoke thought on activity successes and failures. Presenting data in the form of a graph or chart provides a way to combine a lot of data and give the viewer the ability to consider the entirety of the data in a simple way. When done well, a chart or graph may save pages of report space; however, it is important to remember to keep it simple and clear. **Figure 1** depicts two different graphs of the same data, but one of them is clearly misrepresenting the findings. When graphs lack appropriate labels and context, it is easy to read more into the data than what is actually there.

Key features to note when making a graph:

- Describe the chart's data using consistent units.
- Clearly label the axis to avoid an unclear basepoint problem.
- Label the represented data as a bar or point, so that it is clear what it represents.

Errors from Using Percentage Values

Often, converting raw values into percentages is a good way to present data, in particular to show changes. However, the same technique can also promote a serious error in interpreting the data: inconsistent use of terms. Although simple, this error is still often encountered by savvy data handlers. Two terms that are often confused are **percentage point change** and **percentage change**. Percentage points deal with percentages as a unit, so that “1 percentage point = 1 percent.” When percentages are increased, the **percentage point change** is represented by the post-score minus the pre-score (or reversed if the post-test score is lower). On the other hand, if you are describing a change in the values as a fraction or percent of previous data, then you are looking at a change in percentages. **Percentage increase** is post-score minus the pre-score (reversed for percentage decrease), which is then divided by the pre-score and multiplied by 100.

Overreliance on P Values as a Gold Standard

One of the biggest assumptions made in statistics is that by providing a P value associated with a comparative parameter the data is given a “seal of approval.” This subject was important enough to be discussed in a 2014 news article in the prestigious science journal *Nature*¹. In contrast to its modern use as an absolute standard of measure for strength of evidence, the UK statistician Ronald Fisher, who introduced the P value in the 1920s, never meant it to be a definitive test. The original thought was to utilize the methodology solely to determine the probability of an event happening only in the context of the “null hypothesis,” and whether the hypothesis could therefore be disproved based on the available data. The assumptions generated by such a P value are therefore limited, and we may be making a mistake to place too much value when we say “statistically significant.” To emphasize this point, consider the pre-/post-test results testing knowledge before and after a CE activity in the data scenario on the following page.

Believing Non-significance Equals no Effect

In contrast, just because your results are non-significant does not mean there is no effect. There are several reasons you might have obtained non-significant results. One reason may be that the sample size was too small (see previous statistics

Example

Prior to beginning a CE activity, learners scored an average of 40 percent on the pre-test. After the activity was completed the same learners scored an average of 80 percent in the post-test.

This represents a

$80-40 = 40$ **percentage point difference (increase)**,

But a

$80-40/40 * 100 = 100$ **percent increase** in the test scores of these learners.

“Always keep in the back of your mind that, although P values can give an indication of differences, over-reliance on them can prove disastrous!”

series article on sampling); another is perhaps the sample has too much variability. A third reason may be that the effect is small. However, that does not mean it is not important. Small changes can have value. It is important to evaluate the results in the context of your population of learners.

When interpreting studies with non-significance, we can look at the power of the study and the confidence intervals.² Power analyses can help calculate both the minimum sample size required for a study as well as the minimum effect size likely to be detected in a study with a set sample size. Looking at the power analysis, we can then determine if a non-significant result is due to the study being under-powered. In other words, for a non-significant result in a study with low power, we cannot accurately state that the null hypothesis is true. However, if we get a non-significant result with a high-powered study, we can feel more comfortable suggesting that the null hypothesis is true.

Confidence intervals can also be used to interpret non-significant results.² As discussed previously in this series, confidence intervals show the range where the true mean of the population will be found. A non-significant result will have an effect size of zero and the confidence interval will cross zero, suggesting the null hypothesis is true. However, the range, or width, of the confidence interval can give

CONTINUED ON PAGE 6

Data Scenario

You are a provider of CE and decide to test how well a small group of learners has benefited from an educational activity using a comparative pre-/post-test methodology using the same test questions. You obtain the following data: You enter the data for the learners (N = 6) into an Excel document, access the program's functions menu, click "t-test" and select the two data groups. You decide that you want to test to see if the post-test scores are greater than the pre-test (against a null hypothesis that there is no difference), so you select "one-tailed," and because you know that the two groups correspond to data for the same people doing the same test, under "type" you click "paired," Excel gives you the result of **0.003**.

Name	Pre-score /25	Post-score /25
Bob	11	17
Sally	15	19
Steve	19	22
Sue	14	20
Allison	21	19
Mike	16	18
Average Score	15	19.2
Percentage Increase	28%	

This tells you that in every 100 chances; only approximately 0.3 of them would affirm the null hypothesis. This is well below your criteria of **0.05** (5 percent), and consequently you reject the null

The data scenario example illustrates how P values can be misleading; in this case, when the sample size is small, just a few additional sample points can change the interpretation dramatically. If the initial results obtained were presented as representative of a trend in a much larger population, then this could become very embarrassing. Indeed, making premature assumptions with small data samples is one of the most common errors in statistics. However, P values can also be misleading when samples become very large. In this scenario, comparative analysis of pre-/post-test sample groups when N is in the thousands can produce P values that suggest rejection

hypothesis. You mention that your education has been a success, as it improves test scores, and everyone in your office is happy.

Then you unexpectedly receive data from four additional learners, who you originally thought had not submitted their post-tests and consequently were not included in the original calculation. The new data (and new combined average) is shown below:

Name	Pre-score /25	Post-score /25
John	22	17
Sara	21	18
Greg	19	17
Emily	16	16
Average Score	16.9	18.6
Percentage Increase	10.1%	

You notice that although much of the new data does not fall into the same trend as the earlier results, the average of all the data combined is still increased. However, on analyzing the new complete data set (N = 10), using your t-test, you are given a result of **0.11**. As this is above your threshold, you are now forced to conclude the null hypothesis cannot be rejected, and although individual differences can be noted, this is not enough to suggest your education has had a positive impact. Everyone is not so happy!

of the null hypothesis and strong significance, even though the average percentage increase between the groups is actually very small. Does this mean that the change is important? Possibly, but at some point you must determine if the change you have calculated *realistically* suggests the education has been worthwhile. At this point you might feel a little confused as to how to best compare data and to interpret the P value — and rightfully so, but the point is to use appropriate testing and to consider results obtained judiciously. Always keep in the back of your mind that, although P values can give an indication of differences, over-reliance on them can prove disastrous!

“Indeed, making premature assumptions with small data samples is one of the most common errors in statistics.”

more information. If the confidence interval is narrow, this is consistent with the null hypothesis being true and a lack of effect. If the confidence interval is wider, there is more likelihood that there may be a true effect.

Correlation and Causation

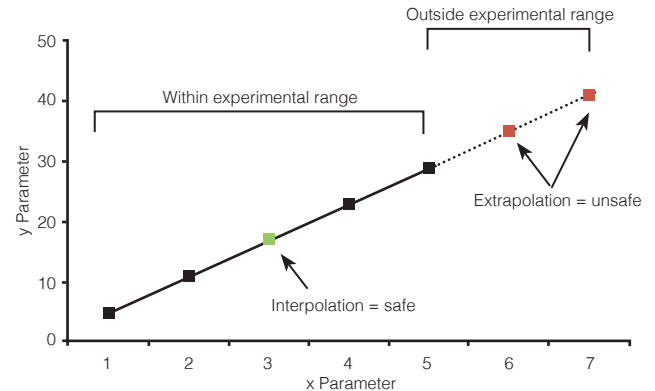
“Correlation does not mean causation” is a common phrase in statistics. Correlation refers to a statistical relationship between two variables. Causation refers to instances where one variable causes another variable to occur. Causation and correlation are examined every day in epidemiology, for example, in cases of food-borne illness. Epidemiologists will investigate all variables that are similar between people who became sick and attempt to determine the cause of the illness. They might find multiple variables in common in a group of people, but realistically only one (e.g. a particular restaurant or food item) will be the cause.

Many people confuse this issue and are very quick to assume that if something precedes an event then it caused it. Most superstitions are based on this assumption — how many black cats have been blamed for a string of bad luck? Another example of this fallacy occurs often in the world of dieting and fitness. A new diet requires ingestion of a particular supplement, reduced consumption of calories and increased physical activity. A person following this diet plan loses weight and assumes it must be due to the supplement. Without an appropriately controlled study to compare Person A with other people who simply dieted and exercised without taking the supplement, it is incorrect to assume causation. For us as CE providers, because of the variability of our learners and the difficulties in making suitable control groups, we should be extra careful in making statements indicating causality.

Extrapolation and Interpolation

Extrapolation is defined as drawing conclusions about a study beyond the range of data (Figure 2) and is another common error in statistics. For example, an inference made about one small sample of learners with a particular trait is applied to all learners who may vary greatly in that trait. Another cause of extrapolation errors involves the use of biased sampling. There are commercials every day stating that four out of five doctors recommend Product X. Extrapolating this result to the whole

Figure 2. Interpolation and Extrapolation to estimate unknown data points.



population would mean that 80 percent of doctors would recommend Product X meaning it must be really good. However, what if only five doctors were used in the sample to generate this statistic or what if the five doctors polled were involved in creating the product? This would be a biased sample. This becomes apparent when a greater number of doctors were polled, in which case, only 15 out of 100 doctors (15 percent) would recommend the product. Doesn't sound as amazing, does it?

Interpolation is a method for determining a new data point within a set of known data points (Figure 2). Similar to extrapolation, it is still a means of estimating a hypothetical value but unlike extrapolation, we can feel safer in our estimate, as we are staying within the experimental range. One of the simplest methods is linear interpolation. When using a formula, the value for an unknown data point can be calculated using the two closest known values on either side of the unknown value and drawing a straight line between them. In Figure 2, the green box represents an unknown value calculated using linear interpolation based on the known values. In addition to Excel and statistics packages, there are numerous online calculators that can be used to calculate linear interpolation. 🧮

Forecast of Next Article

In the final article of this series, Gary Bird, PhD, and Peshia Rubinstein, MPH, CHCP, will provide key takeaways from these articles and list resources for further exploration for the CE professional who is new to the study of statistics.

References and Further Reading

1. Nuzzo, R. 2014. Scientific method: Statistical errors. *Nature*. 506: 150-3.
2. Colegrave and Ruxton. 2002. Confidence intervals are a more useful complement to nonsignificant tests than are power calculations. *Behavioral Ecol.* 14(3). 446-7.

Beginner's Guide to Measuring Educational Outcomes in CEhp



Biostatistics Series Summation and Resources for Continued Skill Building

By Gary Bird, PhD, American Academy of Family Physicians, and Peshia Rubinstein, MPH, CHCP, American Medical Informatics Association

After participating in the entire *Almanac* biostatistics series, the CE professional should be better able to:

- Critique peer-reviewed literature to assess its validity and significance
- Incorporate qualitative and quantitative analytical approaches into the design and planning of CE activities for healthcare professionals

This concluding article in the *Almanac* biostatistics series, which targets the beginning-level learner, concludes as it started: with a restatement of the objectives of the articles.

It is our hope that the series achieved its goals. In this article, we'll summarize the topics the series covered, and then we'll share some evaluation data and recommend additional resources.

Reflection on the series content produces the following definitions and key points:

"Sources of Data in CE," by Lloyd Myers, RPh, and Simone Karp, RPh

- Think backwards from the measure. If you begin with the end in mind, you can work in reverse to make sure that all of your data elements are reasonably within reach.
- Map your activity's desired outcomes to valid data sources. For example, if your activity focuses on a tool that should help primary care practitioners improve suicide risk identification in adults with major depressive disorder, use National Quality Forum's measure description, numerator statement and denominator statement to take the pre- and post-activity measures in this area. Go to www.qualityforum.org/QPS/ and click "Data Source" on the left.

"How to Write Sound Educational Outcomes Questions: A Focus on Knowledge and Competence Assessments," by Erik D. Brady, PhD, CHCP, and Sandra Haas Binford, M.A.Ed.

- Depending on how a CPD professional writes test questions, he or she can measure either a learner's knowledge or competence. By using realistic scenarios, the educator can accurately measure a learner's intent and competence.

“Concepts Involved in Sampling Data,” by Melanie D. Bird, PhD, and Erik D. Brady, PhD, CHCP

- Sampling is the process of selecting participants from a particular population to represent that population as a whole.
- Random sampling is used in CEhp programs, not only to limit bias but for two other tangible reasons: It requires the least amount of forethought in the design of the outcomes tool, and it allows the analyst to report the highest participation possible in the outcomes study.
- The spread of scores around the parameter for our population is called the standard deviation (often abbreviated to SD, or denoted by the Greek letter σ). The spread of scores across the sampling distribution is the standard error (sampling error or SE).
- The lower the standard deviation and the larger the sample size, the smaller the sample error becomes.
- The more diverse the sample is, the larger it will need to be to account for the variability and the more confident we can be in the result.

“Impact of Sampling at Multiple Time Points in Measuring Outcomes of Continuing Education in the Health Professions,” by Gary C. Bird, PhD, and Sandra Haas Binford, M.A.Ed

- As many learners experience information loss after an activity, it is important to gather and report educational outcomes data at longer time points.
- A series of linked CE activities offer opportunities for appropriate and accurate measurement. Outcomes data gathered throughout the series inform us of opportunities to tailor educational events to the evolving needs of learners. Learning occurs only if the content is right and the educational delivery is relevant to the target learner population.

“Basic Concepts of Data Sets,” by Melanie D. Bird, PhD, and Derek T. Dietze, MA, FACEHP, CHCP

- Qualitative data are nonnumeric (words, not numbers) and deemed “categorical,” as they are descriptive in nature, falling into distinct categories such as color, texture, appearance or demographics (sex, type of practice, healthcare specialty). Qualitative data can be described but not measured until they are linked to a numerical scale, which makes them become quantitative.
- Quantitative data are numerical data that can be measured.

“Distribution and Variation in Data Sets,” by Gary C. Bird, PhD; Melanie D. Bird, PhD; and Sandra Haas Binford, MAEd

- A meaningful distribution profile becomes truly apparent when data are graphed. In graphs, data take on a distinctive shape based on the frequency values of the group. A normal distribution, or a bell curve, is most common. However, distributions can also be positively or negatively skewed.
- The distribution of the data determines which statistical test will be most appropriate. For data that follow a normal distribution, parametric tests are used. A parametric test assumes a normal distribution across the population and that the measures are from an equal interval scale. Examples of parametric tests include t-tests and analysis of variance (ANOVA).
- For nominal or ordinal data that may not be distributed normally, nonparametric tests are used. Nonparametric tests are those used to analyze data that have a skewed distribution or when the outcome has limits of detection or outliers. Examples of nonparametric tests include chi-square, the Mann-Whitney test and the Fisher’s exact test.

“How to Analyze Your Baseline, Post-Activity Change Data, Parts 1 and 2: Baseline, Post-Activity Multiple-choice Questions,” by Erik D. Brady, PhD, CHCP, and Derek T. Dietze, MA, FACEHP, CHCP

- A P value (the P means probability) is generated from a test of statistical significance (a mathematical formula). Simply put, the P value represents the role that chance plays in your outcomes.
- In general, a P value of 0.05 or less represents the “gold standard” in scientific research, meaning that 95 percent of the time your findings are statistically significant. This means that there is only a 5 percent likelihood that a calculated change from baseline to post-activity would occur by chance alone if the same education were offered to additional learners of similar demographics.
- A small P value (typically ≤ 0.05) indicates strong evidence that the baseline to post change is real and is not due to chance. A large P value (> 0.05) indicates weak evidence that the baseline to post change is real, and it is more likely due to chance.

“Understanding the Impact of Data and Analysis at the Population Level: How Common Statistical Mistakes Impact Data Interpretation,” by Gary C. Bird, PhD and Melanie D. Bird, PhD

- One of the first things to do as you review data obtained from a CE activity is to ask the question, “Is the data reproducible?” meaning that another person can construct the same study and obtain the same results. Reproducibility is important for generalizing data across a population.
- One of the biggest assumptions made in statistics is that by providing a P value associated with a comparative parameter the data is given a “seal of approval.” However, just because your results are non-significant does not mean there is no effect. Although P values can give an indication of differences, over-reliance on them can prove disastrous!

Evaluation

Since evaluation is part of the cycle that CPD professionals engage in at the conclusion of any activity, we’d like to share in brief the measured outcomes of this series.

The first outcome was a formative qualitative study that the Alliance conducted while gathering Alliance Member Value Statements on this guide. There were five members who chose to comment on the series. All of the feedback was favorable, and we are reprinting two of them here:

- “It was important for me to reach out to you as a long-standing member of the Alliance. I found the statistics series to be journal-worthy! The authors codified complex theoretical constructs and translated these in a manner that supports our collective efforts to integrate these principals in practice. I look forward to the next article.”
- “The new statistical series has added a level of professional value and depth of content that I have shared with my CME planners and certification and instructional design specialists. We recently conducted a staff training utilizing excerpts presented in the July statistics series. Based on feedback from staff, we plan to continue monthly trainings on these topics. Well done!”

A second outcome was the successful submission of a poster to the 2016 World Congress on Continuing Professional Development held in San Diego this past March. The poster focused on the methodology of the series and was organized by Sandra Binford, M.A.Ed. The poster title was “Promoting Adaptive Expertise in Educational Research and Outcomes Analysis Among CEhp Professionals Through a 12-Part Series of Archived Case-Based Newsletter Articles.”

Resources

With the positive evaluation feedback on the series, it seems the *Almanac* has hit upon an important gap in CPD competence. A central concept in the CEhp National Learning Competencies is knowing how to use data for a variety of CPD tasks, from identifying educational gaps to measuring activity success. For those who never studied statistics, the series seems to have opened the door to discussion. For those who studied it a long time ago, the series was a refresher course.

For CPD professionals to continue learning about clinical and educational outcomes measurement, there are many free resources. Search “biostatistics” or “inferential statistics” on the following sites:

- Study.com
- Khanacademy.org
- Coursera.com
- EdX.org
- iTunes University

Massive Open Online Course (MOOC) sites make this – and other – topics of relevance to CPD professionals free and accessible for users. As the number of educational resources can be overwhelming, you may find it helpful to embark on this journey with a mentor or study partner.

Find a Study Partner

If you learn better with real rather than virtual students, get together with a friend or colleague. Establish a set time to study together, enroll in an online course together, get ahold of an introductory book like “Biostatistics for Dummies” or create a syllabus based on the topics in the *Almanac* series. Focus on small increments of material to cover, then practice together by reviewing a journal article that uses the statistical principle you are studying. If you can clearly explain the concept under review to your study partner, your grasp of biostatistics has gotten that much stronger.

And if you and your colleagues complete an entire introductory course or book together, we invite you to write about your experience and its impact on your professional competency in an article for the *Almanac*.

The *Almanac* editors are seeking contributors to create a case-based biostatistics series going forward. Care to share how you measured if an activity made a significant difference in learner outcomes? Email Editor-in-Chief Jacob Coverstone at almanac@acehp.org.