

AMEE GUIDE

'The research compass': An introduction to research in medical education: AMEE Guide No. 56

CHARLOTTE RINGSTED¹, BRIAN HODGES² & ALBERT SCHERPBIER³¹University of Copenhagen and Capital Region, Denmark, ²University of Toronto, Canada, ³University of Maastricht, Netherlands

Abstract

This AMEE Guide offers an introduction to research in medical education. It is intended for those who are contemplating conducting research in medical education but are new to the field. The Guide is structured around the process of transforming ideas and problems into researchable questions, choosing a research approach that is appropriate to the purpose of the study and considering the individual researcher's preferences and the contextual possibilities and constraints. The first section of the Guide addresses the rationale for research in medical education and some of the challenges posed by the complexity of the field. Next is a section on how to move from an idea or problem to a research question by placing a concrete idea or problem within a conceptual, theoretical framework. The following sections are structured around an overview model of approaches to medical education research, 'The research compass'. Core to the model is the conceptual, theoretical framework that is the key to any direction. The compass depicts four main categories of research approaches that can be applied when studying medical education phenomena, 'Explorative studies'; 'Experimental studies'; 'Observational studies'; and 'Translational studies'. Future AMEE Guides in the research series will address these approaches in more detail.

Introduction

This AMEE Guide offers an introduction to research in medical education. It is intended for those who are contemplating conducting research in medical education but are new to the field. This includes those who are generally inexperienced in research as well as those who have previous research experience in the biomedical, but not in the medical education, domain. It is with those readers in mind that we will draw some parallels with research in biomedicine, and indicate where and how research in medical education is similar or different. In addition to some overall principles, the Guide will address the current debate about approaches to medical education research and the study of complex phenomena and interventions. Hence, the Guide we hope may also be of interest to those who already have some experience in research in medical education, and promote future discussion.

The Guide is structured around the process of transforming ideas and problems into researchable questions, choosing a research approach that is appropriate to the purpose of the study and considering the individual researcher's preferences and the contextual possibilities and constraints. The first section of the guide addresses the rationale for doing research in medical education and some of the challenges posed by the complexity of the field. Next is a section on how to move from idea or problem to a research question. This section describes specifically how to place a concrete idea or problem within a

Practice points

- Research in medical education seeks to deepen the knowledge and understanding of learning, teaching and education. It is neither about solving concrete, local problems nor about providing general, universal solutions.
- To get from idea, problem or phenomenon of interest to a research question, it is necessary to have a conceptual, theoretical framework for the study: Identify underlying theories of mechanisms and principles of learning, teaching or education pertaining to the topic and search the literature for 'what is already known' and 'what needs to be investigated further'.
- Research is about taking small steps, making choices and sacrifices in order to focus the topic of inquiry and formulate a general, researchable question.
- There are many research approaches to choose among, each having its own purpose. Four main categories are: Explorative studies aiming at modelling; experimental studies aiming at justifying; observational studies aiming at predicting; and translational studies aiming at implementing.
- The choice of research approach depends on the research question, and often more than one type or mixed approaches are both feasible and necessary.

Correspondence: C. Ringsted, Centre for Clinical Education, University of Copenhagen and Capital Region, Rigshospitalet, 5404, Blegdamsvej 9, 2100 Copenhagen, Denmark. Tel: 45 35455424; fax: 45 35454437; email: charlotte.ringsted@rh.regionh.dk

conceptual, theoretical framework. The following sections are structured around a model that gives an overview of approaches to medical education research, ‘The research compass’. Core to the model is the conceptual, theoretical framework that is the key to any direction. The compass depicts four main categories of research approaches that can be applied when studying medical education phenomena, ‘Explorative studies’, ‘Experimental studies’, ‘Observational studies’, and ‘Translational studies’. Separate sections of this Guide are devoted to general principles of each of the elements of the model. The emphasis is on helping the readers in their search for further information about their topic of interest.

This introduction Guide is the first in a series of AMEE Guides covering Research in Medical Education. Subsequent Guides in the series will address specific research approaches, designs and methods, while other issues will address the topic of conceptual, theoretical frameworks as they relate to research in medical education.

Research in medical education

The altruistic purpose of research in medical education is to deepen the knowledge and understanding of learning and education by studying phenomena, relations and how and why what works for whom. However, one can think of other incentives to engage in research in medical education. The researcher may be driven by genuine scholarly interest in medical education phenomena and/or a wish for getting published in order to gain promotion, fame or a place in the community of medical education scholars. Another purpose could be to justify spending money, time and effort on medical education activities or to attract funding for new educational initiatives or technologies. Finally, taking a scientific approach to innovating medical education practice may be a wise strategy to get academic stakeholders to buy into new ideas and concepts and gain their support for innovations (Grant & Gale 1989). But whatever the incentive may be, there is invariably the need for rigour in the research approach and, in order to get published in scientific journals, clarity as to the

practical relevance of the study and how it contributes new knowledge and understanding of learning and education *in general*.

The challenge

The challenge to the researcher is to place a concrete idea, interest or problem within a general context of learning, teaching and education. Newcomers to the field of medical education research may not be aware of this generalisation perspective and be inclined to focus on local, concrete problems or questions they wish to understand or resolve by collecting and analysing data from evaluation, assessment and audit. For example: ‘How satisfied are participants with the content and format of my course?’, ‘What is the learning outcome measured by a skills test at the end of this class?’, or ‘How well is this programme implemented in our institution as measured by the number of teachers who adhere to the standards?’ But research is not primarily about answering local, concrete questions. In fact, it is the other way round—it is about *researchable problems* relating to general questions about learning, teaching and education that are studied in local contexts. From this perspective, the generalisability of study results is a prime consideration. One aspect of generalisability is how the study contributes general new knowledge about learning, teaching and education. This requires critical appraisal of how the results may have been affected by the context of the study. Another aspect relates to *practical relevance* and how the new knowledge can guide educational practice. However, medical education is a highly complex discipline with huge differences in practices within and across classes, schools, sites, jurisdictions and countries. Adding to the complexity of medical education are the numerous inter-related factors that together constitute the intricate ecology of education (Figure 1).

It would be an unattainable goal for researchers to aspire to finding the one solution or explanation that is uniformly valid in all medical education settings. The most researchers can probably aim for is contributing bits and parts of new knowledge, understanding and reflection on a variety of

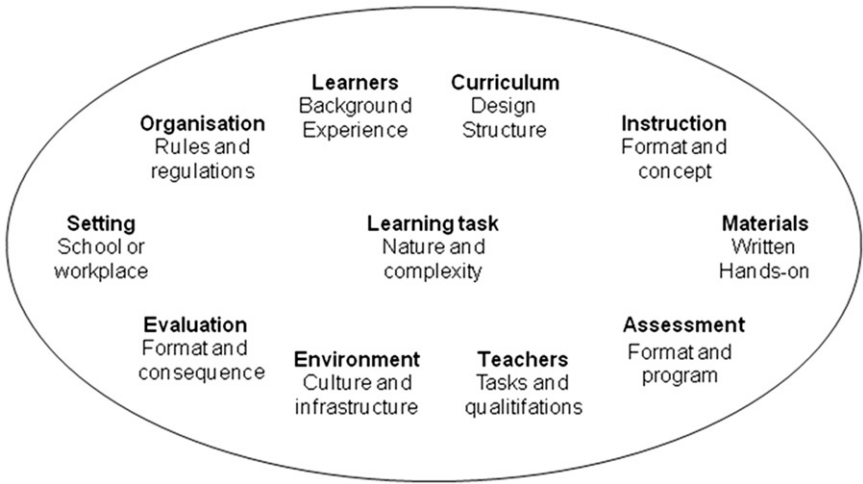


Figure 1. The learning ecology. Numerous aspects interact and contribute to the learning.

phenomena and how and why things work or do not work, and for whom. In a nutshell, research in medical education is neither about solving local problems nor about finding general solutions (Regehr 2010). Researchers would do best to leave those aspects to stakeholders and practitioners, hoping that they will use critical reviews of research findings to better perform their tasks.

An important characteristic of the discipline of medical education is that its practice and delivery are mainly non-standardised. This is rather unlike biomedicine where much biomedical technology, for example drugs or technical procedures, consists of highly standardised interventions. Biomedicine also has the advantage that it has recourse to measurement instruments of exceptional precision in assessing needs and outcomes. Consequently, in biomedicine, generalisations are both possible and relevant to some extent. Nevertheless, even in the biomedical domain, which is heavily influenced by the positivistic philosophies of scientific progress, there is a growing trend towards acknowledging the complexity of practice and broadening the perspectives of research paradigms (Campbell et al. 2000; Ward et al. 2009a; Bunni & Kelly 2010).

From idea or problem to research question

Newcomers to medical education research usually have their scientific interest sparked off by a concrete problem or phenomenon in their local setting. Now, the first step for a researcher is to turn this problem, concern or observation into a *general researchable problem*. This launches them into an endeavour of analysing and identifying which basic concepts are involved in their concrete problem or idea and of inquiring into the underlying mechanisms of learning, teaching, and education that are relevant to their idea. This is quite similar to research in biomedicine, which requires an understanding of the underlying physiology, epidemiology, pharmacology, etc. related to the disease, drug or technology of interest. In other words, the first step in medical education research is situating the idea or problem within a conceptual theoretical framework.

Conceptual, theoretical framework

The conceptual, theoretical framework relevant to a study is a composite of three parts: (1) selecting theories of learning and education that can clarify the underlying mechanisms pertaining to the idea or problem; (2) a critical synthesis of information from the empirical literature identifying what is already known and what is not known about the idea to inform the development of a concrete research topic; and (3) the researcher's individual thoughts and ideas. This framework aids in transforming a personal or local idea or problem into a research problem of general interest. The framework will further assist in formulating a researchable question and choosing an appropriate research approach. In this way, the framework provides a systematic structure and organisation to support the rationale for the study and justify why and how it will be undertaken. Finally, once the study has

been completed, the conceptual, theoretical framework guides the discussion about the generalisability of the results or findings of the study.

The challenge

A literature study is the starting point of the analysis of an idea or problem according to a conceptual, theoretical framework. For those who are new to medical education research, this stage is likely to entail a lot of hard work. However, identifying a framework also saves work, because it helps the researcher to focus on aspects that are worth pursuing and block off irrelevant roads of inquiry. The first step is to identify the overarching themes into which the idea or problem can be subsumed (Figure 1). The next step is to study the basic concepts related to those themes and the related underlying mechanisms of learning, teaching and education. While engaging in this endeavour, the researcher usually discovers that the central idea or problem can be approached from a variety of angles and perspectives. Choices are thus inevitable. In making those choices, the researcher is well advised to choose as the focus of the study aspects for which he or she feels genuine enthusiasm.

Where to look?

There are numerous textbooks on basic concepts of learning, teaching and education, which can be of help in identifying the overarching theme of the idea or problem to be studied. There is however no textbook on conceptual, theoretical frameworks, although there are several articles on the construct (Carpiano & Daley 2006; Bordage 2009; Bunni & Kelly 2010). In addition, some review articles look into a variety of theories as they relate to medical education topics (Mann 2004; Bleakley 2006; Patel et al. 2009a, b; Mann 2011). Good starting points, with plenty of useful references, for a search of what is already known and what is not known but merits further study can be found in a wide array of review papers such as the BEME reviews and AMEE Guides. These papers usually point to areas where further research is needed. BEME guide no. 3, parts 1 and 2, provides help in searching literature databases for medical education (Haig & Dozier 2003a, b), and the AMEE website provides ample links to data resources, www.amee.org.

Formulating research questions

Having identified the topic of the study, the researcher's next step is to formulate a *general* research question or an overall aim of the study, which can then be broken down into more *specific* research questions. The format and wording of research questions are closely linked to the research approach that is chosen. Explorative studies using a qualitative research approach usually ask rather open-ended questions aimed at identifying variables that can explain a phenomenon, whereas experimental studies usually ask closed-type questions relating to pre-defined variables with the aim of justifying a relationship. In this AMEE Guide, the issue of how to formulate

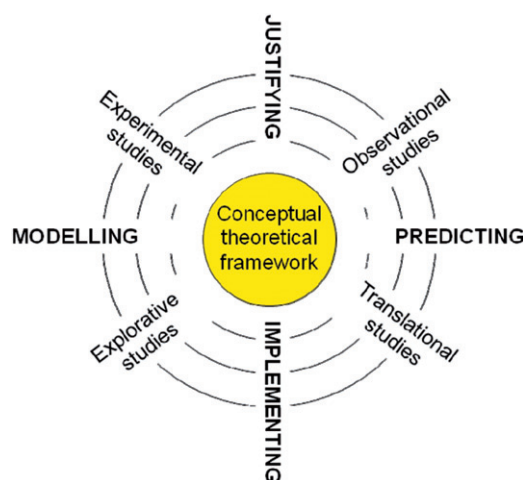


Figure 2. *The Research Compass.* An overview model of approaches to research in medical education.

research questions will therefore be addressed in the individual sections on distinct research approaches.

The research compass

The research compass (Figure 2) is presented as a model that affords an overview of the various approaches to research in medical education. It is a composite of the perspectives presented by Campbell et al. (2000) on approaches to studying complex interventions and Cook et al.'s (2008) framework for classifying the purposes of research in medical education. Several textbooks and articles on research in medical education and biomedicine also inspired the model.

At the centre of the model is the 'conceptual theoretical framework', which is core to any study and the basis of any research approach taken. Applied to the surrounding maze is a compass consisting of four broad categories of studies each relating to different purposes. One category is *explorative studies*, aimed at modelling by seeking to identify and explain elements of phenomena and their relationships. This group of studies includes descriptive studies and studies using qualitative research approaches. Also, under this heading are psychometric studies, aimed at establishing the validity and reliability of measurement instruments. *Modelling* is not only a purpose of explorative studies, but also a precursor to designing experimental studies and defining appropriate interventions and outcomes. The aim of *experimental studies* is *justification* and these studies are typically highly controlled experiments involving homogeneous groups of participants. This type of study, however, is not always feasible in the medical education domain and *observational studies*, i.e. cohort, case-control and associational studies are often better alternatives. Observational studies examine natural or static groups of people and are aimed at *predicting* some sort of outcome. Finally, *translational studies* focus on *implementing* knowledge and findings from research in real life complex settings where people are not alike. This implementation may be followed by the evaluation of both process and outcome, which in turn can lead to the discovery of new unexpected

phenomena, which may prompt further investigation in explorative studies and perhaps new controlled experiments.

We do not claim that 'The research compass' model is exhaustive. Rather, the model seeks to paint the big picture and illustrate how a wide array of research approaches, each in their own right, have a contribution to make to building onto existing knowledge and understanding of learning and education. Unlike a labyrinth, the maze of research has several entries and choices of paths and directions. Moreover, it is not the intention that a 360° endeavour should be run through from start to finish in each research project. Research is about taking small steps by building on prior knowledge and understanding and that can be done at any stages in the maze. Thus, the individual researchers are free to jump on and off at any stage. In other words, they can choose the research approach that fits their needs and preferences, qualitative or quantitative, experimental or translational, as long as the approach is appropriate to the research problem at hand and feasible within the context where the study is to be conducted. Finally, the various research approaches are not exclusive and combinations are often both necessary and feasible. In the following sections, each element of the model will be described in some detail.

Explorative studies and modelling

A mix of rather different research approaches is grouped in the category of explorative studies. Included here are descriptive, qualitative and psychometric studies. The aim of explorative studies is modelling, by identifying, describing and analysing underlying characteristics and mechanisms of phenomena, behaviours, interventions, measurement instruments, etc. Explorative studies ask open-ended research questions such as 'what characterises...?', 'how do people perceive or explain...?', 'which factors can be identified...?', 'what is the validity and reliability of...?'. Based on the results, the researcher synthesises a model or explanation of the topic under investigation.

Modelling also refers to the researcher's preliminary explorations before deciding on a research approach and study design for an experimental, observational or translational study. In this sense, modelling refers to exploring and analysing bits and parts of an intervention and why and how it is expected to work in a particular study. In addition, modelling is used to explore the best method of data collection and selection of study subjects.

Descriptive studies

Descriptions of phenomena, new initiatives or activities, such as curriculum design, instructional methods, assessment formats and evaluation strategies, do not usually qualify as research. This explains why it is increasingly difficult to get this type of study published in scientific medical education journals. However, if the study addresses a research question that relates to a conceptual, theoretical framework, it stands a better chance of being accepted as research. Examples of research questions are 'how do theories of learning and teaching inform the observation of... phenomenon?' or 'how

do theories and prior empirical studies inform the design, enactment or evaluation of...initiative?' Since research is about studying relationships, it is common practice to apply the rather simplistic rule of thumb that for a study to be classified as research it has to be about some sort of comparison or establishing relationships. For descriptive studies, 'comparison' can relate to a conceptual, theoretical framework.

Descriptive studies of medical education used to be quite common in scientific journals and at medical education conferences (Cook et al. 2007; Todres et al. 2007). However, the leaders within the field of research in medical education and the editors of medical education journals have raised the scientific standards for research studies. The problem with simple descriptive studies is that although they may report on a variety of important observations and good initiatives, including evaluation, assessment or audit data, they often lack the generalisation perspective that was discussed in the introductory section of this Guide. There are several well-described methods that can be applied to address this generalisation perspective, including design-based research, action research and case studies. This Guide will not go into detail about these techniques, but ample information can be found in textbooks and journals.

Qualitative studies

In recent years, there has been a rise in interest in using qualitative research methods in health and education research. For many bio-medically trained clinicians, qualitative methods may be less familiar. Pope and Mays (1995) describe the overall goal of qualitative research as the development of concepts that help us to understand social phenomena, in natural (rather than experimental) settings by giving emphasis to the meanings, experiences and views of participants. Because qualitative methods use language-based rather than numerically based data, reading, interpreting and conducting qualitative research requires special methods (Greenhalgh & Taylor 1997).

Theoretically, qualitative research arises from a twentieth century development in philosophy that recognised the importance of language in constructing what we perceive as 'reality'. Many qualitative researchers believe that anything we call 'truth' or 'reality' is filtered through and conditioned by language. Using this 'constructivist paradigm', researchers using qualitative approaches place emphasis on social, historical and individual contexts (Kuper et al. 2008b).

Qualitative approaches are best used for discovering the answers to 'why', 'how' or 'what is the nature of...' type questions. These approaches are used in three primary situations: (1) preliminary to quantitative research; (2) supplemental data (for 'triangulation') with quantitative data; and (3) to explore complex phenomena not amenable to quantitative research. It is important to note that qualitative and quantitative approaches are not opposites. While sometimes portrayed as opposing poles, they are actually different and complementary ways of viewing similar phenomena. For example, imagine a study focusing on diabetes and insulin compliance. A helpful quantitative study might explore the question 'What

is the relationship between non-compliance (as measured by haemoglobin A1C) and disease progression?' A qualitative study might explore the question: 'Why are patients not compliant? How do they view disease progression? What is the nature of their understanding of the concept of "diabetes" and of "compliance"?'

There are many qualitative methods including interviews, focus groups, case studies, text analysis and observation. Researchers should be aware, however, that it is important to understand the conceptual framework that goes with each of these methods. Each belongs to one or more research traditions such as ethnography (Reeves et al. 2008), discourse analysis (Hodges et al. 2008) and grounded theory (Lingard et al. 2008). It is important to understand these 'methodologies' from which the various 'methods' (tools) arise, before using them to collect data.

Obviously statistical analysis cannot be used to analyse language-based data. However, there are many approaches to coding and interpreting qualitative data. Usually this begins by capturing and transcribing data into a textual format for analysis. Kvale summarises five approaches to transcript analysis and coding: meaning condensation, thematic categorisation, construction of narrative, interpretative methods and finally a composite *ad hoc* approach (Kvale 1996).

An important difference from the quantitative paradigm is the way in which 'quality' is understood in qualitative research (Kuper et al. 2008a). Because data are collected in naturalistic settings, giving primacy to the experiences and perspectives of participants, there is no assumption that results will automatically generalise to other situations. Rather, researchers look for evidence that findings are 'transferable' to different contexts. Similarly, the notion of 'reliability' of data is replaced with a concept of 'trustworthiness' and 'validity' with 'authenticity'. The aim of qualitative approaches is to understand perspectives and experiences of participants in their diversity, rather than to reduce data by eliminating statistical 'noise' as in the quantitative tradition. Thus, qualitative researchers speak of 'perspective', including their own perspective, emphasising 'reflexivity' rather than 'bias'. Other elements of quality to attend to are the adequacy of sampling and the 'saturation' of data, authenticity and trustworthiness (good data and good analysis), various forms of 'triangulation' (examining the relationship and fit of data collected in various ways from various sources), 'member checking' (returning data to participants for confirmation), multiple coding and the presence of an 'audit trail' (record of decisions and work done). A rich array of resources (see reference cited above) and courses are available for researchers interested in qualitative approaches.

Psychometric studies

Psychometric studies typically deal with measurement and measurement instruments. Recent decades have seen the publication of numerous studies on assessment, which almost always involve both measurement and measurement instruments. In this section, two notions that are fundamental to measurement are described: validity and reliability. Additionally, attention is given to the development of a measurement or assessment 'instrument'.

Validity

Validity is an important characteristic of a measurement instrument. An instrument is considered to have validity when it has been shown that the instrument does indeed measure what it is purported to measure. In high-stake assessments which are used to inform crucial pass–fail decisions, validity is obviously of paramount importance. Different kinds of validity are distinguished and studied. The four most common types of validity are face, content, criterion and construct validity.

Face validity

In the past, many studies used the concept of ‘face validity’ to indicate that at face value what the instrument measured seemed to correspond with the content of what it was supposed to measure. However, this definition is generally no longer considered acceptable in medical education research. Journal editors are unlikely to accept studies in which this rather naive definition is used. They require more substantial evidence of validity (see later). However, there is an exception to this rule. In the literature on simulation in medical education, it is accepted practice for researchers to use the term ‘face validity’ to indicate that in the judgment of experts the model provides an acceptable approximation of the reality of medical practice.

Content validity

The main element in content validity is sampling. For example, in constructing a test that students must pass at the end of an educational unit test, designers have to make sure that the questions or items of the test are an appropriate representation of the educational content of the unit. In order to ensure the content validity of such a test, it is common practice to provide a blueprint describing the different content areas covered by the unit, and to determine the number of test questions/items to be included for each area.

Content validity is also an important aspect of the development of a questionnaire to measure a certain construct. In developing such a questionnaire, investigators should start with an extensive review of the theory underlying the construct and, ideally, they should also look for other published questionnaires. At a certain point in the development process, a number of themes and questions have been identified. At that stage, it is customary and it helps for content validity to invite experts to give their opinion on the relationship between the underlying theory and the instrument.

Unfortunately, some colleagues continue to entertain rather simplistic ideas about the development of tests and questionnaires. It is important to realise that in fact it is a highly complicated and challenging task to make it and to be sure that there is content validity. In research practice, investigators often need to construct a questionnaire to gather data about a topic on which little knowledge is available. In such cases, the best line of action is to first conduct a qualitative study in order to gain a comprehensive picture of the idea or construct

involved and in that way get information about the content. It is impossible to say ‘how much’ content validity is achieved. In practice, there are many papers in which this aspect of validity does not receive a great deal of attention.

Criterion validity

Criterion validity depends on the amount of agreement between the outcomes of the test or measurement of interest and those of another test or assessment. The other test or assessment is referred to as the criterion. Two types of criterion validity are distinguished: predictive validity and concurrent validity. When the predictive validity of a measurement is to be determined, the criterion measurement occurs at a later time. This is for example the case when the objective is to establish the extent to which high school grades are predictive of results in medical school. Some well-known examples from medical education can be found in Swanson et al. (1991) and Ramsey et al. (1989).

When concurrent validity is at stake, the criterion is sampled at the same time as the measurement of interest.

Construct validity

This type of validity is the most difficult to determine. Typically, investigators try to establish correlations between scores, measurements and performances that are all assumed to be related to a particular theory or construct. A typical example is a study in which investigators compare the scores of residents and consultants on a certain measurement. A difference between consultants and residents is interpreted as an argument in support of construct validity, because of their differences in expertise are expected to impact their performance. However, often this conclusion is too simple, since there are numerous differences between residents and consultants that can influence measurement outcomes in different ways.

Streiner and Norman (2008) describe three methods that are often used to establish construct validity: extreme group comparisons; convergent and discriminant validity studies and multi-trait multi-method studies. The interested reader can read more about this topic in their book.

Reliability

Two types of reliability are distinguished: reproducibility and internal consistency. Both refer to the consistency of measurement outcomes. Reproducibility of a measurement means that a student who gains a high score on a certain test will achieve the same high score when the test is administered on a subsequent occasion. Internal consistency implies that test items or questions all measure the same subject or construct.

Reproducibility

The test–retest method and the equivalent method are commonly used to investigate reproducibility.

With the test–retest method, the same test is repeated after a certain period of time. The correlation between the two scores

is estimated by calculating the reliability coefficient, which has to reach a certain level. If the time interval between the two measurements is too short, the investigator should take account of the possibility of a learning effect due to the first measurement. If too much time has elapsed between the two measurements, the coefficient will be lower, because the test results may be subject to influences from many other variables.

The equivalent method requires the use of two different instruments to measure outcomes during the same period. The instruments must be comparable. Because it is very difficult to develop two instruments that are truly comparable, this method is less commonly used than the test-retest method.

In the assessment literature two different types of reproducibility are distinguished: intra-rater and inter-rater reliability. When one or more raters are involved in judging a candidate, intra-rater reliability is determined by the answer to the question: *are the judgments produced by the same rater consistent over time?* In other words, will the same rater give the same judgment on a candidate after a certain period of time (intra-rater reliability)? The question to be asked in relation to inter-rater reliability is: *do different raters give the same judgment on the same candidate?*

Internal consistency

The internal consistency of a test or measurement depends on the degree to which the items in a scale are homogenous in relation to each other. In other words, to what extent do the items measure the same underlying construct or concept? There are two ways to determine internal consistency: calculating the reliability coefficient or splitting the test and calculating the coefficient for both parts.

The most frequently used statistic that is calculated to assess internal consistency is Cronbach's alpha. For dichotomous data the Kuder-Richardson coefficient is more appropriate. It should be noted that these coefficients are not stable characteristics of a certain instrument. They are related to a specific measurement at a certain moment, and therefore subject to change. Another way to examine the internal consistency of a test is to split the test in two halves and calculate if both halves yield the same results. This can be done using the Spearman-Brown formula.

The development of an instrument

The process of instrument development starts with a clear description and definition of the topic under investigation and of what is to be measured and how. This requires an extensive literature search which serves several objectives:

- (1) To look for theoretical models or theories that are related to the topic and can be used to underpin the instrument;
- (2) To find out what other investigators have studied in relation to the topic;
- (3) To search for existing instruments and ascertain whether these are valid and reliable.

For the new investigator, it can be very helpful to first discuss plans for developing an instrument with more

experienced investigators, who can recommend which papers are important to read and who are familiar with the work that has already been done by others. Common advice from an experienced investigator will be to choose another approach than developing a new instrument.

After the topic has been clearly described and a clear definition formulated, and when the investigator has decided to develop a new instrument, the next step is to define the content to be measured. Depending on the topic, there are several ways to define content and construct a blueprint. For example, if the instrument is developed to provide information about a certain procedure, an analysis of the procedure could be the starting point for defining content. At this stage, it may turn out that there is not enough knowledge available about the topic to define content. When that happens, it is often advisable to first conduct a qualitative investigation, such as a focus group or interview study.

Expert meetings are often used to determine the content of the measurement. Currently, theoretical background is an important consideration in this phase of the development of an instrument. Theory can be useful in decisions about content, because it can inform choices concerning the viewpoint or perspective that the investigator will take in studying the topic.

Once the content has been defined, the instrument can be composed by the investigator. It is wise to invite colleagues to critically review the instrument avoid redundancy, check the wording, etc.

When the first draft of the instrument is completed, the time has come to try out the instrument. It is important to find out if the 'users' understand the questions and how much time it takes for them to complete the instrument. In larger research projects, the pilot phase usually involves a larger number of participants in order to enable statistical analyses to be conducted. After improving the instrument based on the results of the pilot phase, the instrument is ready for use (Box 1).

In order to determine the type of statistical analysis that is appropriate for the data collected with the instrument, it is important to consider the kind of scale that is used. Is it a nominal, ordinal, interval or ratio scale?

A nominal scale is a scale where a number is assigned to the variables by the investigator. For example: 1 = male and 2 = female. These numbers neither denote a numerical value nor a certain order. They merely serve to categorise the data. In an ordinal scale, information is ranked in a certain order, but this does not imply that the differences between the scores are equal. In an interval scale, the distances between scale points are all equal. For example, the distance between 10 and 20 kilograms is the same as that between 40 and 50 kg. There are

Box 1. Development of an instrument.

1. Literature search
2. Definition of the topic for the investigation
3. Decision about the content
4. Composition of the instrument
5. Pilot testing and improving the instrument
6. Using the instrument

many situations in which it is impossible to determine with certainty whether or not the differences between scale points are equal. A ratio scale is an interval scale with a true zero value, such as height or weight. In practice, investigators often assume that an interval scale is used, because this type of scale allows a greater variety of statistical analyses.

Experimental studies

The goal of experimental studies is to provide justification. In other words, they seek evidence of the effects of an intervention. The research questions are typically of the closed ended type: ‘Does this intervention work?’ or ‘Is this intervention better or more efficient than another one?’ To make sure that the effect is due to the study intervention, the researcher has to control for numerous factors that might influence the result. These so-called confounders relate to characteristics of the population studied, the intervention itself or the context of the study. By carefully controlling the experiment, the researcher seeks to minimise the undesirable effects of confounders. There are several ways to achieve this. First, by selecting a homogeneous study sample, second by standardising the intervention and the measurement instrument, and third by having a control group that does not receive the intervention. In the following examples of experimental study designs, general comments will be made about specific issues and challenges related to experimental research in medical education. However, this Guide only provides some overall principles and readers are referred to the literature for details on how to conduct experimental studies including statistical considerations and analyses.

There are many types of experimental studies. In biomedicine, the Randomised Controlled Trial (RCT) is generally considered the pinnacle of research designs, which is superior to all other designs in providing valid evidence of the effect of an intervention. An RCT investigates whether an intervention works under strictly controlled and standardised circumstances. Typically, it is anticipated that the use of strict inclusion and exclusion criteria and random allocation to the intervention or control group will ensure that confounders related to individual characteristics will be equal in both groups, leaving the intervention the only difference between them. The classical RCT compares the intervention to no intervention (placebo) (Table 1). Participants are randomised to either the intervention or the control group. A baseline measurement is performed in both groups followed by the intervention in one group and a placebo in the other group, and finally the measurement is repeated in both groups. Subsequently, the data are analysed for differences within and between the two groups in the expectation that the experiment

will demonstrate a change in the intervention group, but not in the control group.

In medical education, research there are several problems with the classic RCT design. First, in education, a comparison between intervention and no intervention is inherently meaningless. For instance, if the intervention is a course on resuscitation skills, the group who receives the intervention is almost bound to do better on a later performance measurement than a control group who did not attend the course simply because there is no natural course of development of resuscitation skills. Hence, in medical education, it makes more sense to compare a new intervention to an appropriate alternative rather than to no intervention.

The second problem with the design illustrated in Table 1 is that in education the baseline measurement may influence the result either through test-enhanced learning, motivation to learn or by stimulating recall of prior knowledge. Moreover, if the same test is used before and after the intervention, there is a risk that participants will remember the test. Clearly, pre-tests are to be used with caution in medical education research. However, this faces the researcher with a dilemma, because without a pre-test how does one estimate the amount of learning that has occurred.

Third, the timing of the measurement after the intervention can be crucial to the results. When a learning outcome is measured immediately after an intervention, the results can be quite misleading as evidence of the occurrence of true learning. True learning is characterised by a relatively sustainable change of capacity in relation to whatever aspect of learning is studied—knowledge, skills, attitudes, motivation, emotions and socialisation. In order to ascertain the occurrence of sustainable change, it is recommended to conduct follow-up measurements either by measuring retention of learning some time after the intervention and/or by measuring whether participants are able to apply the newly acquired capability to other but similar tasks or contexts, in other words, by measuring transfer of learning.

A four-group RCT design that takes account of these three problems is depicted in Table 2. The participants are randomised to four groups. The top two include a before measurement and the bottom two do not. The table also indicates that the intervention is compared to an appropriate alternative and that some sort of follow-up measurement is planned.

The four-group RCT design is an ideal that can be difficult to attain in practice. One feasible alternative is to break up the study into a pilot study measuring learning outcome before

Table 1. The randomised controlled Trial (RCT).

Classic RCT				
Randomisation	I _{Group}	M _{Before}	I	M _{After}
	C _{Group}	M _{Before}	C _{Placebo}	M _{After}

Note: I = intervention, C = control and M = measurement.

Table 2. The four-group RCT design with follow-up.

Solomon design		Follow-up				
Randomisation	I _{Group 1}	M _{Before}	I _{New}	M _{After}	M _{Retention}	M _{Transfer}
	C _{Group 1}	M _{Before}	I _{Alternative}	M _{After}	M _{Retention}	M _{Transfer}
	I _{Group 2}		I _{New}	M _{After}	M _{Retention}	M _{Transfer}
	C _{Group 2}		I _{Alternative}	M _{After}	M _{Retention}	M _{Transfer}

Note: I = intervention, C = control, and M = measurement.

and after the intervention, and a main study before measurements but including follow-up measurements.

Finally, in medical education, randomisation can be difficult to accomplish and often the researcher has to settle for other means of establishing groups for comparison. These non-randomised studies are also called quasi-experimental studies.

The challenge

Irrespective of the design, RCT or quasi-experimental, experimental studies entail some challenges that need to be addressed before a specific research question can be formulated. These challenges pertain to six inter-related elements: the setting, the study population, the intervention and control conditions, the outcome measure, and the time of measurement. These elements can be abbreviated to SPICOT (Haynes 2006), and they must all be included in the specific research question of an experimental study. Although it is commonly said that the research question drives the study, identifying and defining the SPICOT elements is an iterative process of modelling in accordance with the chosen conceptual, theoretical framework while constantly taking account of the inter-relatedness of the SPICOT elements. Adjusting one element inevitably affects one or more of the other ones. Moreover, selecting the SPICOT elements is also affected by practical constraints in the context of the study and some choices and sacrifices will be inescapable. Unfortunately, there is no such thing as a perfect experimental study. However, rigour in clarifying and describing each of the elements and the choices made is crucial to the quality of the study, to enable conclusions to be drawn from the results and to getting published.

How to define the SPICOT elements

First and foremost, the intervention and the alternative treatment or control conditions must be defined carefully. This is a tough task in medical education, since an intervention, an instruction or a course usually include several elements that each impact on learning in various ways. Furthermore, there is no general agreement on terminology in medical education. The concepts 'problem-based learning' or 'conventional didactic teaching' may mean different things in different contexts. Careful and accurate description is of the essence. Hypothesising why and how an intervention might work can be helpful. Equally important is the opposite question—why might an intervention not work? These issues should be dealt with during the problem formulation before the actual study is undertaken.

After the intervention and control conditions have been defined, the next step is to identify or develop a valid outcome measure and decide on the time of the measurement. The challenges of these issues were described in the sections on psychometric studies and problems with RCTs.

Finally, an appropriate study sample should be defined. Here, appropriateness relates to both quantity and quality. As for quantity, the number of participants depends on the size of the expected effect, the sensitivity of the measurement instrument and the variance among the participants. Power

calculation can be used to estimate sample size. The reader is referred to statistical texts for further information on this topic, but the general rule is that the larger the expected effect, the higher the sensitivity of the instrument and the lower the variance in the study sample, the smaller the size of the sample that is required to allow conclusions to be drawn based on the results. In experimental studies, it is common to define strict inclusion and exclusion criteria in order to achieve a sample that is as homogeneous as possible. However, similarity in the sample can be a threat to the generalisability of results to the actual people for whom the intervention is intended in the real world. In other words, it can detract from the external validity of the study. The following example illustrates this phenomenon. It is not uncommon for experimental studies on the effect of advanced virtual reality simulators to be conducted on samples of medical students. Such samples have the advantage of being quite homogeneous regarding prior learning for they are all novices to the advanced procedure to be trained on the simulator. However, principles of learning and instruction that are derived from a study among novices might well be quite different and even contrary to those that are suitable for advanced or experienced learners (Wulf & Shea 2002; Haynes 2006). This threat relates to the cognitive, emotional as well as social aspects of learning (Illeris 2004). Another problem is the measurement instrument, which may be appropriate for measuring the initial stages of learning, but fail to reflect more advanced stages.

Observational studies

Central to observational studies is the prediction of some sort of relationship between variables. These studies differ from experimental studies in several ways. The participants are natural or static groups of people rather than selected samples, although usually these studies also have some sort of inclusion and exclusion criteria. In observational studies, participants are not allocated to a group by the researcher, but groups are determined based on for instance classes within a school or year of entry in the school.

There are several reasons for choosing an observational study approach over an experimental one. An RCT may not be feasible, because of difficulties in administering disparate treatments for the experimental and control group, opposition to an RCT from students/authorities who have strong emotional beliefs regarding the benefits/drawbacks of a proposed intervention, or difficulties recruiting participants within a reasonable period of time. For example, recruitment can be a problem in postgraduate medical education programmes, where the number of trainees in one specialty is generally rather low compared to the number of undergraduate medical students in 1 year. Second, experiments are very useful to demonstrate *if* an intervention works in a highly selected group of people under controlled circumstances, but RCTs often have poor external validity, i.e. the results do not predict if the intervention will actually work in less homogeneous groups in educational practice. Finally, observational studies are preferable when the research question has a long-term perspective and also takes account of unintended effects.

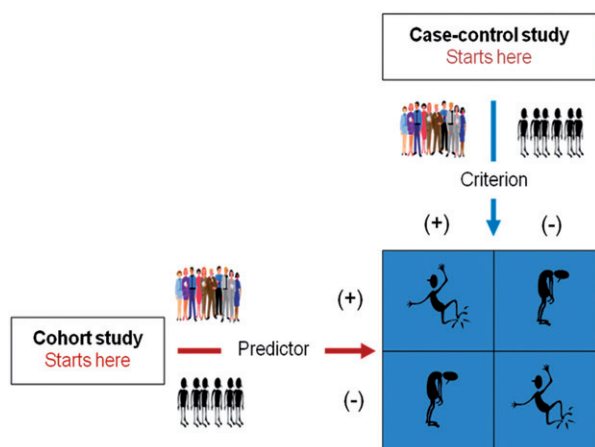


Figure 3. The difference between Cohort studies and Case-control studies is the direction of inquiry.

Types of observational studies

Three categories of observational studies will be addressed in brief in this Guide: cohort studies, case-control studies and associational studies. The difference between cohort and case-control studies is the *direction of the inquiry*, Figure 3. In cohort studies, the inquiry starts with the predictor variable, i.e. an exposure or characteristic that is different between cohorts, whereas in case-control studies, the inquiry starts with the criterion variable, i.e. an outcome that is different between groups. The researcher is 'happy' if the majority of subjects studied match the hypothesis of relationship between prediction and outcome criteria, i.e. falls into the upper left and the lower right quadrants in the figure. The researcher is 'sad' when subjects fall into the upper right or lower left quadrants as that is in opposition to the hypothesis of a relation between the predictor and the outcome criteria.

Cohort studies

In cohort studies, *the inquiry* starts with the predictor variable, i.e. a cohort's exposure to a particular educational event or a cohort's specific characteristics, and investigates how the predictor variable is related to a criterion variable (the outcome). The research question could be: 'Does attending a medical school with a problem-based learning (PBL) curriculum predict future competence as a doctor?' (Norman et al. 2008). The predictor is exposure to PBL and the criterion is future competence measured by peer assessment of performance as a doctor in practice. Cohorts attending PBL schools are compared to cohorts attending non-PBL schools and thus the predictor variable defines the groups. Incidentally, a retrospective cohort study on this topic was unable to establish such a relationship (Norman et al. 2008).

Cohort studies can also follow a cohort over time and monitor outcomes on pre-defined criterion variables. For example, data on exam results by the end of medical school (the predictor here is a characteristic rather than an exposure) can be analysed to determine an association with data on the quality of health care (criterion variables) provided by the

same cohort as practising physicians (Tamblyn et al. 2002; Cadieux et al. 2011).

Case-control studies

In case-control studies, *the inquiry* starts with the criterion variable and investigates exposure or characteristics of participants in relation to that outcome (case group) compared to a matched group of participants (control group). Thus, the criterion variable defines the groups. For example, the criterion could be disciplinary action by a Medical Board and the predictor could be ratings of professionalism during medical school (Papadakis et al. 2004). In this particular study, the case group included physicians who had been subject to disciplinary action, while the control group consisted of physicians who had not faced similar action. The groups were matched for medical school, year of graduation and choice of specialty. For each group, data on professionalism during medical school were analysed and compared. Incidentally, this study was able to establish such a relationship (Papadakis et al. 2004).

Figure 4 depicts the cohort study of Norman et al. and the case-control study of Papadakis et al. The cohort study could not demonstrate a relation between having attended a PBL school and future assessment as 'excellent' or 'concern'. There was an equal percentage of study subjects from each cohort assessed as 'excellent' and an equal percentage assessed from the two cohorts assessed as 'concern'. The case-control study demonstrated that the prevalence of the predictor 'concern regarding professional behaviour in medical school' was twice as high in the case-group compared to the control-group, indicating an association between the predictor and the criterion, i.e. disciplinary action by a Medical Board.

Case-control studies and retrospective cohort studies are faster to perform than prospective cohort follow-up studies. Case-control studies are especially useful when the criterion variable is categorical (yes/no, present/absent), the outcome is of rare occurrence and there is a long time delay until the outcome is manifest. However, case-control studies are considered inferior to prospective cohort studies in providing evidence of causal relationships. Nonetheless, a case-control study can be a good starting point for exploring the association between a set of criterion and predictor variables, which can subsequently be investigated with stronger research designs.

Associational studies

Associational studies do not necessarily compare groups. An example is cross-sectional studies, which provide a snapshot of certain variables in a variety of people and investigate how they are associated, Figure 5.

The predictor could be clinical experience measured in number of years and the criterion could be the quality of skill performance measured by a test. In the theoretical example in the left-hand box in Figure 5, the correlation coefficient of 0.86 indicates a good relationship. With perfect correlation it would be 1.00. Associational studies can be contrasted with experimental studies. In associational studies, different participants are analysed, and there are no groups. The variables are called

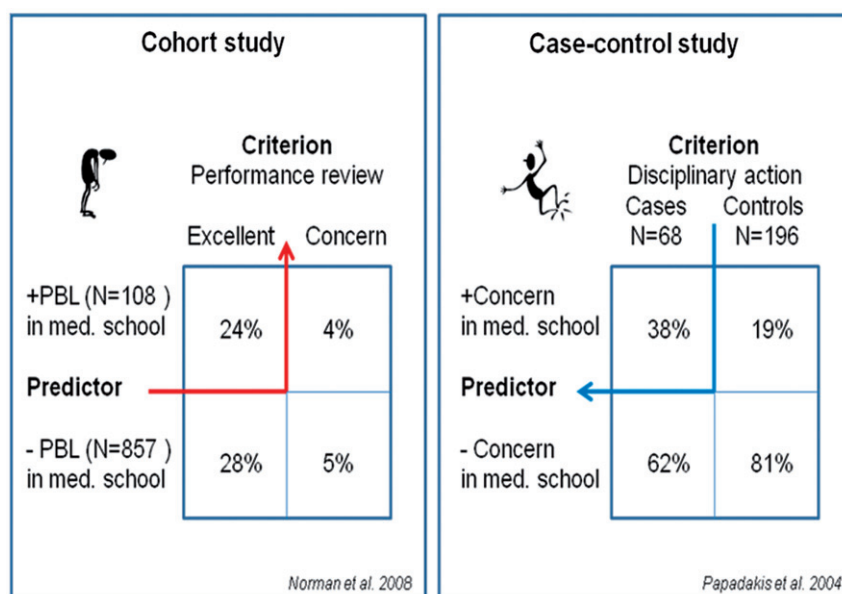


Figure 4. Examples of a cohort and a case-control study.

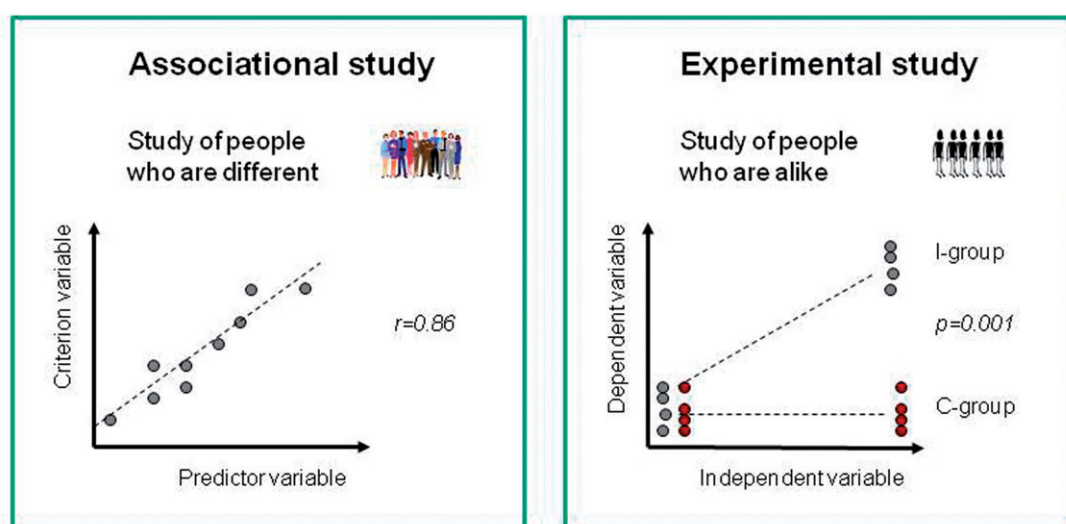


Figure 5. The difference between associational and experimental studies. I-Group = Intervention group and C-Group = Control group. The independent variable is the cause/the intervention and the dependent variable the effect/outcome.

the predictor and the criterion variables. In experimental studies, care is taken to recruit participants that are as alike as possible. They are then allocated to two groups that either receive an intervention (I-group) or serve as control group (C-group) that does not receive the intervention and finally the outcome is measured. The variables are called the independent variable (the cause) and the dependent variable (the effect or outcome). A real effect of the intervention is considered to be present when a difference in outcome between the groups is demonstrated with a p -value indicating the probability that this difference occurred by chance.

In observational studies, there is often more than one predictor variable and hence a need for more sophisticated methods to analyse the contribution of each of these to the outcome, i.e. multivariate regression analyses. Cohort and

case-control studies are therefore frequently accompanied by associational analyses of variance (Teherani et al. 2005; Norman et al. 2008; Cadieux et al. 2011). It is outside the scope of this introductory Guide to go into this, but future AMEE Guides in the research series will address these research approaches more extensively.

The challenge

The challenge of observational studies is that the researcher cannot control for bias due to confounding variables. Another challenge is the choice of methods for analysing and interpreting the data (Colliver et al. 2007). That is why it is highly advisable for researchers to consult a statistician as soon as they begin to design the study. Finally, although some sort

of relationship may be established by an observational study, this can only be interpreted as an indication and not as evidence of a *causal* relationship, in part because of the numerous known and unknown factors that may influence the results. Based on a sound theoretical background, researchers should provide a carefully considered explanation of why a relation is expected, while in-depth knowledge of data management and interpretation is a crucial prerequisite to conduct this type of study. It should be noted, however, that observed associations, while not to be taken as straightforward evidence, can be used to direct further investigations into potentially causal relationship using an RCT or other controlled experimental study. Finally, the kind of data that are needed to establish associations between educational processes, outcomes and quality of health care requires comprehensive databases containing clinical as well as educational data. The absence of such data repositories has prompted a call for educational databases that prospectively collect trainee information over time and across multiple institutions for the purpose of doing observational studies (Cook et al. 2010).

Translational studies

Translational research is a concept that arose in the biomedical domain based on a concern that scientific discoveries failed to be translated into clinical practice where they can be of benefit to patients and the public (Zerhouni 2005). The road from the basic science lab to clinical practice is long and in essence comprises three major steps. The first is transferring basic science discoveries to clinical patient-oriented studies. The second is synthesising knowledge from empirical clinical studies and producing clinical practice guidelines. This part is also called knowledge translation (Graham et al. 2006). The third step is implementing the new knowledge in clinical practice and evaluating whether, how and why it works in a variety of complex real life settings, i.e. efficiency studies. However, in contrast to the view of the translational process as a unidirectional trajectory from bench to bedside, there is increasing awareness of a cyclic or multi-directional inter-relatedness of knowledge transfer and research between basic and applied clinical science (Ward et al. 2009a; Rubio et al. 2010).

Translational research has relevance to medical education in more than one respect. First of all, research can examine how medical education practice can support the implementation of the concepts of translational research as they apply to biomedicine. This can be done by including these issues in formal undergraduate and postgraduate medical curricula. Another aspect relates to strategies of development and implementation of best evidence clinical guidelines. Here, research in the arena of continuous medical education and professional development of practitioners can make an important contribution by revealing principles of how medical professionals look for and assimilate new knowledge and eventually use it to change their behaviour and practices (Slotnick 1999; Davis et al. 2008). Third, translational research in the medical education domain can be interpreted as research that investigates how learning in the classroom or simulated settings is transferred to competence in real life

practice, which eventually leads to better patient outcomes (McGaghie 2010). And finally, translational research in medical education can be viewed as research to elucidate how principles from the basic and applied sciences of learning and instruction can be translated into everyday medical education practice.

Translational studies in medical education are discussed further under the next three headings: knowledge creation, knowledge implementation and efficiency studies.

Knowledge creation (reviews)

Knowledge creation studies are investigations of prior research. They are known as reviews, and highly comprehensive examples include systematic, realist and BEME reviews (BEME=Best Evidence Medical Education). Just like any other research approach, a systematic review must be situated within a conceptual, theoretical framework; start from a research question(s), use a systematic approach to collect data and use well defined inclusion and exclusion criteria and a systematic approach to data analysis and interpretation.

Systematic reviews are typically quantitative in orientation and seek to compile evidence of the effect of a certain intervention in order to create a summary effect size. There are quite a few systematic reviews on medical education published in medical education journals as well as biomedical journals. However, quantitative reviews are complicated by the variance across studies with regard to the SPICOT elements, described in the section on experimental studies. Differences in SPICOT elements across studies, if unattended, may lead to collation of studies with outcomes in opposite directions. An example is the choice of measurement points, which can have a crucial impact on the results. The task of taking these problems into account is difficult for the reviewer, because SPICOT elements are poorly described in many experimental studies. It is therefore not surprising that many systematic quantitative reviews demonstrate the small overall effect sizes or are inconclusive. Nevertheless, systematic reviews are invaluable to the individual researcher who is contemplating a study of a certain topic because they offer an excellent overview of prior research as well as a wealth of useful references.

The problems of systematic quantitative reviews lie in part in insufficient attention to conceptual, theoretical frameworks. This is a major focus in realist reviews, which draw upon quantitative empirical evidence, placing it in a conceptual, theoretical framework. This kind of reviews is looking for theories to explain differences in empirical results across studies (Pawson et al. 2005; McGaghie 2010; Wong et al. 2010).

The BEME reviews are a mix of systematic quantitative reviews and realist reviews. They often have a primarily practical focus and provide guidance on the best evidence available and how that can inform educational practice.

Less comprehensive are critical, narrative reviews, which rather than claiming to be exhaustive seek to determine the essence of prior research and identify problems that merit further study. Conducting a narrative review is typical when starting a research project and is used for setting the scene and

describing the conceptual, theoretical framework for the study (Eva 2008).

Knowledge implementation

At the start of this section on translational research, it was pointed out that implementation of clinical guidelines and adoption of evidence-based medical practice are highly topical issues in biomedicine. Clinical guidelines are based on systematic reviews and considered a key tool for knowledge implementation in medicine. The counterpart in medical education is not so straightforward, in part because of the absence of evidence-based guidelines and in part because many medical educators are largely unaware of the medical education literature. AMEE Guides take a pragmatic perspective and strive to stimulate critical thinking and reflection, while providing practical advice and support. One might say that they are evidence inspired rather than evidence based. The BEME reviews show more similarity to the format of clinical guidelines in the medical domain. However, the complexity of medical education and the interaction of numerous factors (Figure 1) render it virtually impossible to create general guidelines to fit any contextual environment, given the huge variety in local characteristics, possibilities and constraints (Regehr 2010).

Many educators experience frustration over the prevailing traditional teaching modes in their institutions and work hard to implement and sustain innovative educational practices. One way to distribute knowledge about best evidence medical education is by running courses, workshops and seminars. This is a widespread strategy, but despite some evidence of a beneficial effect of faculty development courses and initiatives, little is known about why and how they work or why they sometimes do not work (Mann 2004; Steinert et al. 2006, 2009). Another way to promote change is to join with administrators and stakeholders in the development of guidelines rather than taking a directive and instructional approach (Grol & Grimshaw 2003). Involving stakeholders in the process is one of a recently validated set of quality indicators for developing and implementing clinical guidelines (Brouwers et al. 2010). A large part of these indicators can be used for the development of medical education guidelines as well. Moreover, arrays of theoretical, conceptual frameworks relating to the dissemination of knowledge are also applicable to studies on knowledge translation (Estabrooks et al. 2006; Ward et al. 2009a). Finally, turning educational innovations into research projects may be another promising strategy in the medical world (Grant & Gale 1989). One reason is the openness signalled by the research question and another is the involvement of stakeholders in the endeavour. More will be said about this issue in the next section.

Efficiency studies

While experimental studies address questions of whether interventions work under controlled conditions, efficiency studies seek answers to the question of what works how and for whom in real-life settings. In biomedicine, there is increasing awareness that multiple factors influence the

adoption of knowledge from basic science in practice settings and also that a multi-angle approach to evaluating interventions is needed (Craig et al. 2008; Ward et al. 2009b). These phenomena are the province of efficiency studies, which investigate complex interventions rather than simple or single-type treatments. Complexity relates to the number of components and flexibility of the intervention, the variety of behaviours required from users, the number of groups or organisational levels involved and the variance in outcomes (Craig et al. 2008). Usually, efficiency studies are very practical in perspective and seek answers that are meaningful to key stakeholders and relevant to decision making (Tunis et al. 2010). These studies include a variety of outcome and process measures in order to evaluate effects on individuals, systems and organisations, including cost-effectiveness. In medical education, the concept of evaluating effects on several levels, i.e. reaction, learning, behaviour and organisation, is well-known from Kirkpatrick's (1998) model. However, the new trend in efficiency studies is to broaden the perspectives of studies from mere evaluation to issues of development, feasibility and implementation. Development includes identifying evidence and underlying theory and modelling processes and outcomes. The feasibility part includes piloting procedures, considering recruitment and retention and determining effect sizes of a large-scale main study. Implementation issues comprise dissemination, surveillance and monitoring and long-term follow-up. Conducting the four parts of efficiency studies—development, feasibility, implementation, and evaluation – is viewed as an iterative back-and-forth process rather than a staged model (Craig et al. 2008). Further details of the framework can be found at the Medical Research Council web-site, www.mrc.ac.uk/complexinterventionsguidance.

Although the counterpart to developing and evaluating complex interventions in medical education research could be quite similar to that in biomedicine, there will inevitably be many areas in medical education that do not fit the variety of research approaches recommended by MRC guidelines. However, the principles and rigour that are recommended are equally relevant to medical education and areas like cost-effectiveness for example deserve more attention in medical education research. Several of the concepts of efficiency studies are quite similar to those of design-based research with its key components of 'design, enactment, implementation and evaluation'. Both efficiency studies and design-based studies can be quite substantial in scope. Both approaches are similar in their recommendation of multi-site involvement, which implies involvement of numerous participants in the work. However, parts of the results that emerge during the endeavour can inform the understanding of what works why, or why not, and for whom, and can be published as individual papers. The conceptual, theoretical framework, the research questions and the rigour of the approach are the characteristics that distinguish efficiency studies from mere evaluation, assessment and audit—as mentioned in the introduction of this guide. So far, the literature provides more reports on research proposals for efficiency studies than actual study results.

Concluding remarks

This AMEE Guide is intended as an introduction to research in medical education. In addition to outlining some basic principles of research within the medical education domain, it presents an overview of the variety of research approaches that can be applied. After reading this introduction, the reader should not be under the impression that applied and theory-driven research are opposite poles, but rather have gained insight into how they are intertwined and mutually interdependent in furthering the shared goal of extending the knowledge and understanding of medical education. Future AMEE Guides will address the various research approaches in considerably greater detail.

Declaration of interest: None of the authors have any conflicting interest regarding this article.

Notes on contributors

CHARLOTTE RINGSTED MD, MHPE, PhD, is a Professor of Medical Education and Director of Centre for Clinical Education, University of Copenhagen, Faculty of Health Science and Capital Region of Denmark.

BRIAN HODGES, MD, PhD, is a Professor, Scientist and Director of The Wilson Centre for Research in Education, University of Toronto, Faculty of Medicine and Vice President, Education, University Health Network, Canada.

ALBERT SCHERPBIER, MD, PhD, is a Professor of Medical Education, Scientific Director of Institute for Education, Faculty of Health, Medicine, and Life Sciences, and Associate Dean of Education Maastricht University, The Netherlands.

References

- Bleakley A. 2006. Broadening conceptions of learning in medical education: The message from teamworking. *Med Educ* 40:150–157.
- Bordage G. 2009. Conceptual frameworks to illuminate and magnify. *Med Educ* 43:312–319.
- Brouwers MC, Kho ME, Browman GP, Burgers JS, Cluzeau F, Feder G, Fervers B, Graham ID, Hanna SE, Makarski J. 2010. Development of the AGREE II, part 1: Performance, usefulness and areas for improvement. *CMAJ* 182:1045–1052.
- Bunniss S, Kelly DR. 2010. Research paradigms in medical education research. *Med Educ* 44:358–366.
- Cadieux G, Abrahamowicz M, Dauphinee D, Tamblyn R. 2011. Are physicians with better clinical skills on licensing examinations less likely to prescribe antibiotics for viral respiratory infections in ambulatory care settings? *Med Care* 49:156–165.
- Campbell M, Fitzpatrick R, Haines A, Kinmonth AL, Sandercock P, Spiegelhalter D, Tyrer P. 2000. Framework for design and evaluation of complex interventions to improve health. *BMJ* 321:694–696.
- Carpiano RM, Daley DM. 2006. A guide and glossary on post-positivist theory building for population health. *J Epidemiol Commun Health* 60:564–570.
- Colliver JA, Markwell SJ, Verhulst SJ, Robbs RS. 2007. The prognostic value of documented unprofessional behavior in medical school records for predicting and preventing subsequent medical board disciplinary action: The Papadakis studies revisited. *Teach Learn Med* 19:213–215.
- Cook DA, Andriole DA, Durning SJ, Roberts NK, Triola MM. 2010. Longitudinal research databases in medical education: Facilitating the study of educational outcomes over time and across institutions. *Acad Med* 85:1340–1346.
- Cook DA, Beckman TJ, Bordage G. 2007. A systematic review of titles and abstracts of experimental studies in medical education: Many informative elements missing. *Med Educ* 41:1074–1081.

- Cook DA, Bordage G, Schmidt HG. 2008. Description, justification and clarification: A framework for classifying the purposes of research in medical education. *Med Educ* 42:128–133.
- Craig P, Dieppe P, Macintyre S, Michie S, Nazareth I, Petticrew M. 2008. Developing and evaluating complex interventions: The new Medical Research Council guidance. *BMJ* 337:a1655.
- Davis N, Davis D, Bloch R. 2008. Continuing medical education: AMEE Education Guide No. 35. *Med Teach* 30:652–666.
- Estabrooks CA, Thompson DS, Lovely JJ, Hofmeyer A. 2006. A guide to knowledge translation theory. *J Contin Educ Health Prof* 26:25–36.
- Eva KW. 2008. On the limits of systematicity. *Med Educ* 42:852–853.
- Graham ID, Logan J, Harrison MB, Straus SE, Tetroe J, Caswell W, Robinson N. 2006. Lost in knowledge translation: Time for a map? *J Contin Educ Health Prof* 26:13–24.
- Grant J, Gale R. 1989. Changing medical education. *Med Educ* 23:252–257.
- Greenhalgh T, Taylor R. 1997. Papers that go beyond numbers (qualitative research). *BMJ* 315:740–743.
- Grol R, Grimshaw J. 2003. From best evidence to best practice: Effective implementation of change in patients' care. *Lancet* 362:1225–1230.
- Haig A, Dozier M. 2003a. BEME Guide no 3: Systematic searching for evidence in medical education—Part 1: Sources of information. *Med Teach* 25:352–363.
- Haig A, Dozier M. 2003b. BEME Guide no. 3: Systematic searching for evidence in medical education—Part 2: Constructing searches. *Med Teach* 25:463–484.
- Haynes R. 2006. Forming research questions. *J Clin Epidemiol* 59:881–886.
- Hodges BD, Kuper A, Reeves S. 2008. Discourse analysis. *BMJ* 337:a879.
- Illeris K. 2004. The three dimensions of learning. Copenhagen: Roskilde University Press/Leicester: NIACE.
- Kirkpatrick D. 1998. Evaluating training programs. San Francisco (CA): Berrett-Kochler.
- Kuper A, Lingard L, Levinson W. 2008a. Critically appraising qualitative research. *BMJ* 337:a1035.
- Kuper A, Reeves S, Levinson W. 2008b. An introduction to reading and appraising qualitative research. *BMJ* 337:a288.
- Kvale S. 1996. *InterViews: An introduction to qualitative research interviewing*. Thousand Oaks (CA): Sage Publications.
- Lingard L, Albert M, Levinson W. 2008. Grounded theory, mixed methods, and action research. *BMJ* 337:a567.
- Mann KV. 2004. The role of educational theory in continuing medical education: Has it helped us? *J Contin Educ Health Prof* 24(Suppl 1):S22–S30.
- Mann KV. 2011. Theoretical perspectives in medical education: Past experience and future possibilities. *Med Educ* 45:60–68.
- McGaghie WC. 2010. Medical education research as translational science. *Sci Transl Med* 2:19cm8.
- Norman GR, Wenghofer E, Klass D. 2008. Predicting doctor performance outcomes of curriculum interventions: Problem-based learning and continuing competence. *Med Educ* 42:794–799.
- Papadakis MA, Hodgson CS, Teherani A, Kohatsu ND. 2004. Unprofessional behavior in medical school is associated with subsequent disciplinary action by a state medical board. *Acad Med* 79:244–249.
- Patel VL, Yoskowitz NA, Arocha JF. 2009a. Towards effective evaluation and reform in medical education: A cognitive and learning sciences perspective. *Adv Health Sci Educ Theory Pract* 14:791–812.
- Patel VL, Yoskowitz NA, Arocha JF, Shortliffe EH. 2009b. Cognitive and learning sciences in biomedical and health instructional design: A review with lessons for biomedical informatics education. *J Biomed Inform* 42:176–197.
- Pawson R, Greenhalgh T, Harvey G, Walshe K. 2005. Realist review—A new method of systematic review designed for complex policy interventions. *J Health Serv Res Policy* 10(Suppl 1):21–34.
- Pope C, Mays N. 1995. Reaching the parts other methods cannot reach: An introduction to qualitative methods in health and health services research. *BMJ* 311:42–45.
- Ramsey PG, Carline JD, Inui TS, Larson EB, LoGerfo JP, Wnrich MD. 1989. Predictive validity of certification by the American Board of Internal Medicine. *Ann Int Med* 110:719–726.
- Reeves S, Kuper A, Hodges BD. 2008. Qualitative research methodologies: Ethnography. *BMJ* 337:a1020.

- Regehr G. 2010. It's NOT rocket science: Rethinking our metaphors for research in health professions education. *Med Educ* 44:31–39.
- Rubio DM, Schoenbaum EE, Lee LS, Schteingart DE, Marantz PR, Anderson KE, Platt LD, Baez A, Esposito K. 2010. Defining translational research: Implications for training. *Acad Med* 85:470–475.
- Slotnick HB. 1999. How doctors learn: Physicians' self-directed learning episodes. *Acad Med* 74:1106–1117.
- Steinert Y, Mann K, Centeno A, Dolmans D, Spencer J, Gelula M, Prideaux D. 2006. A systematic review of faculty development initiatives designed to improve teaching effectiveness in medical education: BEME Guide No. 8. *Med Teach* 28:497–526.
- Steinert Y, McLeod PJ, Boillat M, Meterissian S, Elizov M, Macdonald ME. 2009. Faculty development: A 'field of dreams'? *Med Educ* 43:42–49.
- Streiner DL, Norman GR. 2008. *Health Measurement Scales: A practical guide to their development and use*. 4th ed. Oxford: Oxford University Press.
- Swanson DB, Case SM, Nungester RJ. 1991. Validity of NBME Part I and Part II in prediction of Part III performance. *Acad Med* 66:S7–9.
- Tamblyn R, Abrahamowicz M, Dauphinee WD, Hanley JA, Norcini J, Girard N, Grand'Maison P, Brailovsky C. 2002. Association between licensure examination scores and practice in primary care. *JAMA* 288:3019–3026.
- Teherani A, Hodgson CS, Banach M, Papadakis MA. 2005. Domains of unprofessional behavior during medical school associated with future disciplinary action by a state medical board. *Acad Med* 80:S17–S20.
- Todres M, Stephenson A, Jones R. 2007. Medical education research remains the poor relation. *BMJ* 335:333–335.
- Tunis SR, Benner J, McClellan M. 2010. Comparative effectiveness research: Policy context, methods development and research infrastructure. *Stat Med* 29:1963–1976.
- Ward V, House A, Hamer S. 2009a. Developing a framework for transferring knowledge into action: A thematic analysis of the literature. *J Health Serv Res Policy* 14:156–164.
- Ward VL, House AO, Hamer S. 2009b. Knowledge brokering: Exploring the process of transferring knowledge into action. *BMC Health Serv Res* 9:12.
- Wong G, Greenhalgh T, Pawson R. 2010. Internet-based medical education: A realist review of what works, for whom and in what circumstances. *BMC Med Educ* 10:12.
- Wulf G, Shea CH. 2002. Principles derived from the study of simple skills do not generalize to complex skill learning. *Psychon Bull Rev* 9:185–211.
- Zerhouni EA. 2005. Translational and clinical science—Time for a new vision. *N Engl J Med* 353:1621–1623.