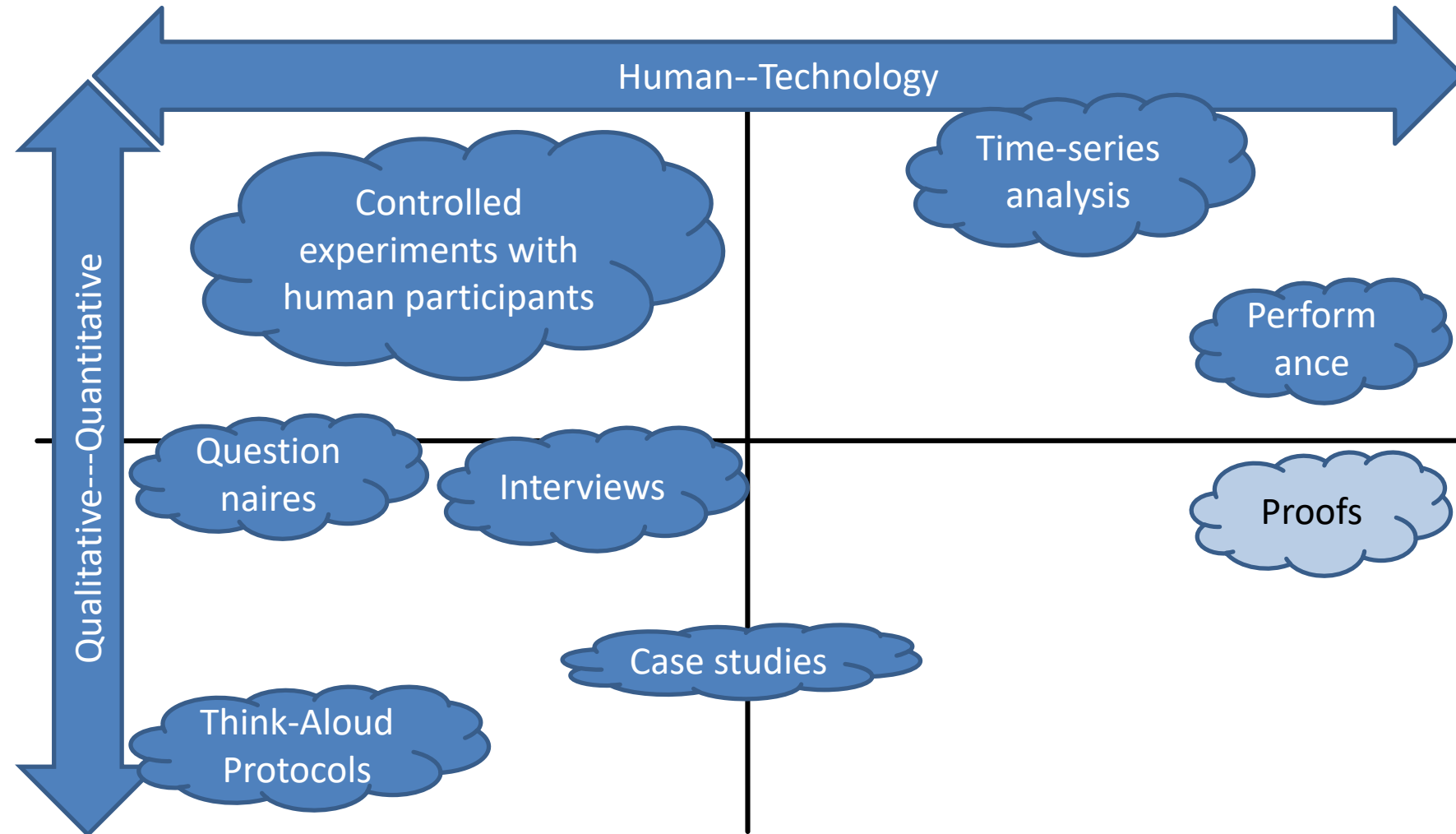


Controlled Experiments

Assignment

- Read excerpts of the following papers:
 - How Do Professional Developers Comprehend Software? (Section II, skim Section III)
 - An Experiment About Static and Dynamic Type Systems (Section 4, skim Section 5)
- What do you think of the experiment
 - What would you do in the same way? Why?
 - What would you do differently? Why?

Overview



Learning Goals

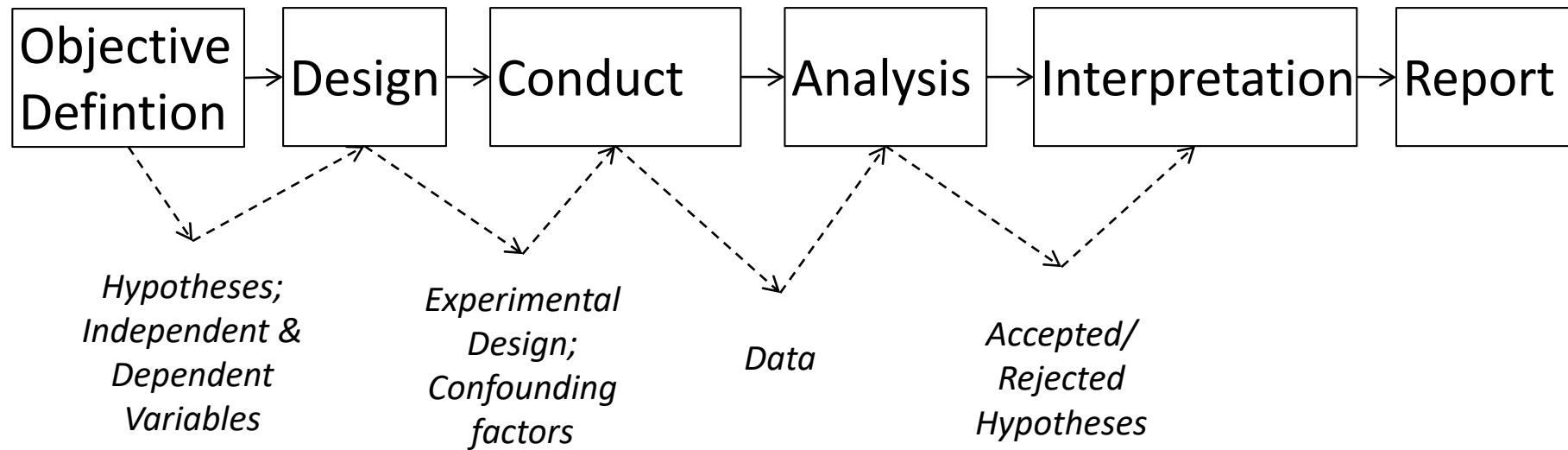
- Design good research hypotheses
- Design an experiment with high internal validity and high external validity



Controlled Experiment: Definition

- Systematic study
- One or more factors are varied
- Everything else is held constant
- Result of systematic variation is observed

Experimental Phases





Variables

Independent Variables

- Varied by experimenter on purpose and systematic
- Also called factor or predictor (-variable)
- Has alternatives, levels, or treatments
- Examples:
 - Programming language or paradigm
 - User interfaces
 - Interaction methods

Dependent Variable

- Result of an experiment
- Depends on variation of independent variable
- Is observed
- Examples:
 - Productivity of programmer
 - Bugs in a program
 - Operator error

Latent Variables

- Construct
- Not observable directly
- Examples
 - Program comprehension
 - Intelligence
 - Mental model of user

Operationalization

- Defining operations that allows you to measure variables
- Must not contradict common sense
- Example:
 - Program comprehension
 - Number of bugs in a program
 - Development time

Task

- Find operational definition for the following variables:
 - Usability of new UI
 - Maintainability of a program



Hypotheses

- Expectations of results
- Expectations need to be justified, e.g., in theory or practice
- Hypotheses need to be simple and clear
- Hypotheses need to be falsifiable
- Falsifiability (Suggested reading: Make yourself familiar with the term and explain its role for experimental design)

Hypotheses - Bad Examples

- Bad source-code comments are bad for program comprehension
- Good source-code comments are good for program comprehension

How can we do better?

- Comments that describe every statement of source code do not affect the time developers need to understand a source-code snippet
- Comments that contain wrong information about source code increase the time developers need to understand a source-code snippet
- Comments describing the purpose of source-code statements decrease the time developers need to understand a source-code snippet

Why do we need Hypotheses?

- They guide us when designing an experiment
- Prohibit *Fishing for Results*
- Connects theory and empirical research
 - Derived from theory
 - Evaluated with empirical research

Task

- State a hypothesis on the following research questions:
 - Does Python increase productivity of developers?
 - Is Python better than C++?
 - Is the new UI more productive?
 - Keep in mind that the hypothesis needs to be evaluated, justified, and the variables operationalized
- = Example for an exam question



Design

Validity

- Do we measure what we want to measure?

Internal Validity

- Amount to which the value of the dependent variable can be explained by the systematic variation of the independent variable

External Validity

- Amount to which results are transferable to other circumstances (participants, material,...)
- Generalizability

Homework Assignment

- Research other kinds of validity

Threats to Validity

- Confounding factors:
 - Influence dependent variable in addition to independent variables
 - Learning effect
 - Hawthorne effect
 - Mono-method bias
 - Selection bias

Task

- Evaluating the effect of new UI: What confounding factors exist?
- How could the influence of these factors be controlled?



Confounding Factors

- There are numerous confounding factors
- Carefully identify and control their influence:
 - Randomization
 - Matching/parallelization/balancing
 - Define confounding factors as independent variable
 - Keep confounding factor constant
 - Analyze afterwards

Randomization

- Random number generator
- Toss a coin
- Throw a dice
- ...
- Issues:
 - Groups need to be sufficiently large
 - 5 per group seems to be too low, 10 seems to be sufficient

Matching/parallelization/balancing

Participant	Value
P5	65
P9	56
P3	42
P4	34
P10	24
P6	23
P7	21
P8	16
P2	12
P1	5

Group A	Group B
65	56
34	42
24	23
16	21
12	6

odd-even-even-odd/
ABBA

Matching/parallelization/balancing

- Drawback compared to randomization
 - Confounding factor needs to be measured
 - How to measure programming experience? How to measure intelligence?
- Advantage compared to randomization:
 - More detailed knowledge about parameter

Define confounding factors as independent variable

- Is varied systematically by experimenter
- Confounding factor is operationalized
- Experience with tool:
 - Haskell/low experience
 - Java/much experience
 - Jave/low experience
 - Haskells/much experience

Doing the Math...

- 23 confounding factors, each with 2 levels
= 8 388 608 possible combinations
- How many participants do I need to cover each combination
 - at least 10 participants per group
 - 83 886 080 (i.e., Germany)

Keep confounding factor constant

- Only one level of confounding factor
- Programming experience:
 - Only undergraduate students
 - Only programming experts
- Intelligence
 - Only students with a certain grade

Analyze afterwards

- Measure factor during experiment
- Analyze influence of a variable after the experiment
- Issues
 - Could show that results are useless

Relation between Internal and External Validity

- Both kinds of validity request different things:
 - Internal: control everything
 - Extern: general setting
- And now?
 - First maximize internal validity
 - Then increase external validity step by step

Quality Criteria of Empirical Studies

- Validity
- Reliability
- Objectivity

Quality Criteria of Empirical Studies

- Reliability:
 - Accuracy of measurement instrument
- Objectivity:
 - Execution of experiment must not depend on person of experimenter
 - The same experiment, conducted by a different experimenter, should produce the same result

Example

- Scale for measuring the weight:
 - Valid
 - Reliable depending on quality
 - Digital scale is more objective, as everyone sees the same number (analogous leaves more wiggle room)
- The same scale for measuring the height
 - Less valid
 - Reliable depending on quality

Further Kinds of Validity

- Construct validity
 - Describes, how well the construct is being measured
 - E.g., program comprehension, intelligence, performance
- Statistical Conclusion Validity
 - Suitability of statistical methods
- Face validity, convergent/discriminant validity, criterion validity, content validity... [wikipedia]



Experimental Designs

Designs

- Between vs. Within Subject
= With vs. without repeated measures
- One-factorial vs. multi-factorial
= One vs. several independent variables
- Univariate vs. Multivariate
= One vs. several dependent variables

Why Experimental Designs?

- Instruction to act
- Makes communication easier
- Decision for statistical analysis

How to Select a Design

- Depends on:
 - Effect size
 - Sample size
 - -> The bigger both are, the less the influence of confounding factors manifest
 - -> When both are small, a suitable design is very important
- Unfortunately, you will very often have small samples and unknown effect sizes

One-Factorial

Between-Subjects

- Participants are divided into two groups
- As many group as there are levels of the independent variable
- Results are compared between groups

Group	Levels
A	Textual Annotation
B	Background colors

```

21 public class PhotoListScreen extends List {
22
23     //Add the core application commands always
24     public static final Command viewCommand = new Com
25     public static final Command addCommand = new Com
26     public static final Command deleteCommand = new C
27     public static final Command backCommand = new Com
28
29     public static final Command editLabelCommand = ne
30
31     // #ifdef includeCountViews
32     public static final Command sortCommand = new Com
33     // #endif
34
35     // #ifdef includeFavourites
36     public static final Command favoriteCommand = new
37     public static final Command viewFavoritesCommand
38     // #endif
39
40     /**
41     * Constructor

```

```

21 public class PhotoListScreen extends List {
22
23     //Add the core application commands always
24     public static final Command viewCommand = new
25     public static final Command addCommand = new C
26     public static final Command deleteCommand = ne
27     public static final Command backCommand = new
28
29     public static final Command editLabelCommand =
30
31     public static final Command sortCommand = new
32
33     public static final Command favoriteCommand =
34     public static final Command viewFavoritesComme
35
36     /**
37     * Constructor
38     */
39     public PhotoListScreen() {
40         super("Choose Items", Choice.IMPLICIT);
41     }

```

Issues

- Variance between participants (i.e., inter individual differences) can be large
- -> 10x (What does 10x Mean? Measuring Variations in Programmer Productivity. Steve McConnell.)
- Sufficient number of participants
- Balancing between groups

Within-Subjects

- Inter individual differences need to be controlled for
- Each participant is exposed to all levels of an independent variable

One Group	Session 1	Session 2
	Background colors	Textual Annotations

Issues

- Learning effects
 - Especially with creative tasks
 - You need different, but similar tasks at the same time
- Ordering effects
- Intra individual differences
 - Fatigue
 - Motivation
- Mortality

Crossover

- Each participant is exposed to all levels
- Comparison between and within groups is possible

Group	Session 1	Session 2
A	Background colors	Textual annotations
B	Textual annotations	Background colors

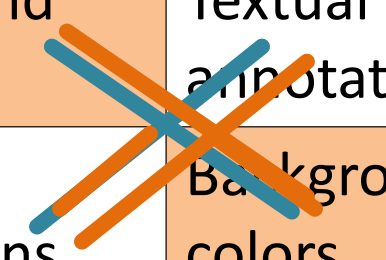
Issues

- Intra individual differences
- Inter individuelle differences
- Mortality

Benefits

- Check for learning effects:
 - Difference between both sessions for both levels
- Check for ordering effects:
 - Difference between both sessions for one level

Gruppe	Session 1	Session 2
A	Background colors	Textual annotations
B	Textual annotations	Background colors



Comparison

Property	Between-Subjects	Within-Subjects	Cross-Over
Sample size	2	1	2
Group balancing	2	1	2
Learning effects	2	3	1
Ordering effects	2	3	1
Mortality	1	2	2
Motivation, fatigue	1	2	2
Experiment duration	1	2	2
Internal validity	2	2	1
External validity	2	2	1

Multi Factorial Designs

Latin Square

Group	Task 1	Task 2
A	Background colors	Textual annotations
B	Textual annotations	Background colors

- Special case of cross over
- But different task in sessions -> Task is second factor

Two-factorial, Between-subjects

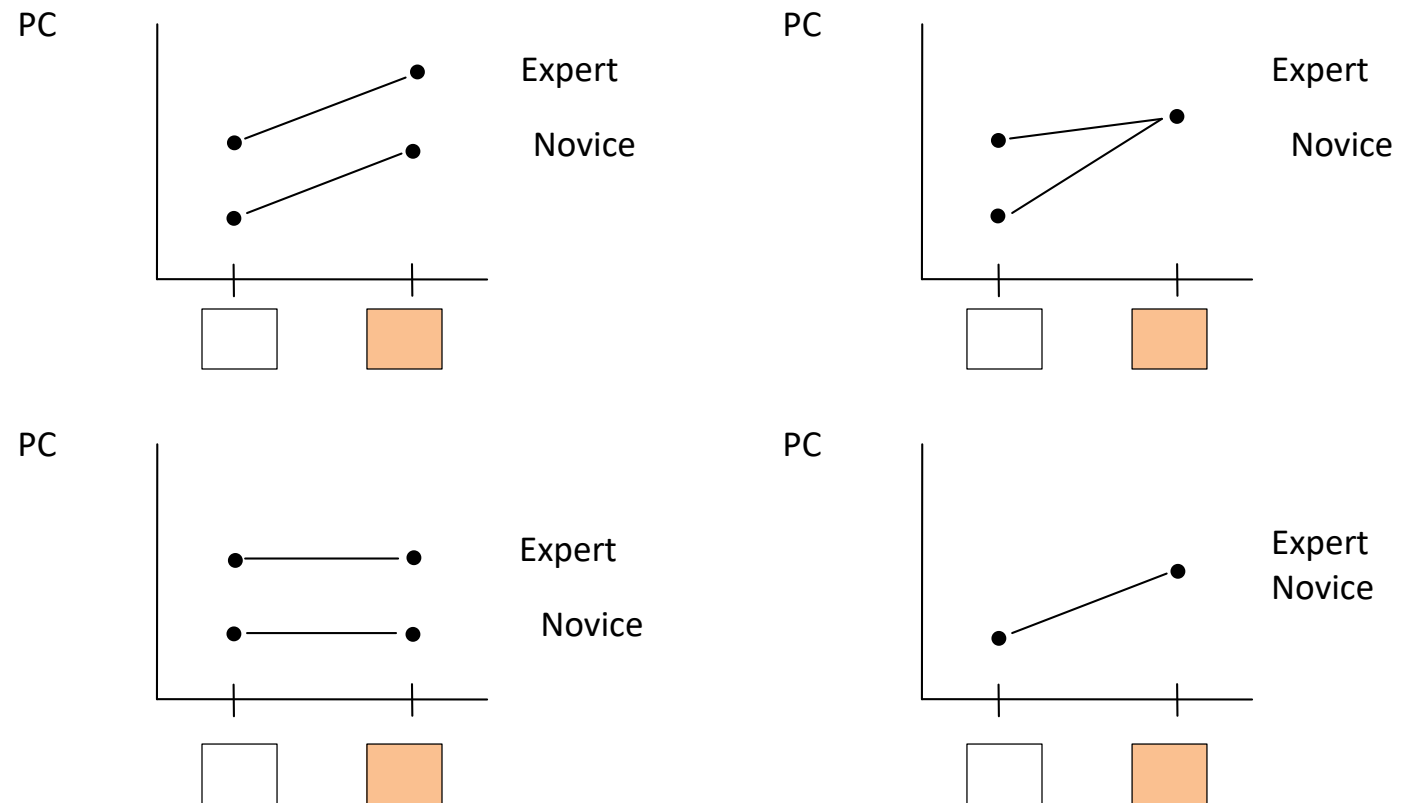
- Programming experience

Variables	Groups
Background color/ novice	Group A
Background color/ expert	Group B
Textual/ novice	Group C
Textual/ Expert	Group D

Two-factorial, Within-Subjects

Group	Session 1	Session 2	Session 3	Session 4
Background color/ novice	Group A	Group D	Group C	Group B
Background color/ expert	Group B	Group A	Group D	Group C
Textual/ novice	Group C	Group B	Group A	Group D
Textual/ Expert	Group D	Group C	Group B	Group A

Main- and Interaction Effects



Multi-factorial Designs

- In case the shown designs are not sufficient
- 4-factorial design (2x2x3x2)
- Higher-order interaction

		C ₁		C ₂		C ₃	
		B ₁	B ₂	B ₁	B ₂	B ₁	B ₂
A ₁	D ₁						
	D ₂						
A ₂	D ₁						
	D ₂						

Selecting a Design

- Prefer a simple as possible design
- Carefully consider benefits and disadvantages
- Consider resource constraints



Conduct

What Can Go Wrong?

- Everything!
- Pilot studies:
 - Test material and tools
 - Check that data is actually stored
 - Check instructions for participants
 - ...
- **Exactly** tell participants what they need to do
- Observe that participants do **exactly** what you told them they should do
- Do a warm-up task, so that participants can familiarize with everything

Ethics

- Effort for participants need to be in line with insights that you gain
 - Evaluating teaching methods
 - Evaluating medicine
- Ensure anonymity of participants
- Be nice to your participants, they voluntarily invest their time

Learning Goals

- Design good research hypotheses
- Design an experiment with high internal validity and high external validity



Task

- Following statements:
 - Program in Java can be easier debugged
 - Learning to program is easiest with Haskell
 - Novice programmers should start with object orientation
 - ...
- Define:
 - Hypotheses
 - Dependent and independent variables and operationalization
 - Confounding factors and how to control them
 - Experimental design

Literature

- Jutta Markgraf, Hans-Peter Musahl, Friedrich Wilkening, Karin Wilkening, and Viktor Sarris. *Studieneinheit Versuchsplanung*, 2001. FIM-Psychologie Modellversuch, Universität Erlangen-Nürnberg.
- Natalia Juristo and Ana Moreno. *Basics of Software Engineering Experimentation*. Kluwer, 2001.
- Claes Wohlin. *Experimentation in Software Engineering*. Springer, 2000.
- William Shadish, Thomas Cook, and Donald Campbell. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Houghton Mifflin Company, 2002.
- James Goodwin. *Research in Psychology: Methods and Design*. Wiley Publishing, Inc., 1999.
- Steve Easterbrook, Janice Singer, Margaret-Anne Storey, and Daniela Damian. *Selecting Empirical Methods for Software Engineering Research*. In *Guide to Advanced Empirical Software Engineering*, pages 285–311. Springer, 2008.
- Steve McConnell. *What does 10x Mean?* In *Making Software*, O'Reilly, 2010.
- Urban Wiesing. *Die Ethik-Kommissionen – Neuere Entwicklungen und Richtlinien*. Deutscher Ärzte-Verlag, 2003.