Design, Durchführung und Präsentation von Experimenten in der Softwaretechnik

Inhalt

1. Zusammenfassung der Papers
2. Fehler in Design, Durchführung und Präsentation
3. Richtlinien für „saubere“ Experimente

Grundlage ist ein Paper in den „IEEE Transactions on Software Engineering“ (2002) über ein Experiment, daß die Verbesserung der Programmqualität durch die Anwendung formaler Methoden zeigen soll.

Versuchspersonen waren dabei Informatik-Studenten im dritten Jahr (vermutlich Bachelor), welche in zwei Gruppen geteilt wurden.

CONTROL GROUP (CG): 13 Teams zu je 2 Studenten

FORMAL METHOD GROUP (FMG): 6 Teams zu je 2 Studenten

Unterschied: die FMG-Mitglieder absolvierten zusätzlich zu gemeinsamen Lehrveranstaltungen Vorlesungen über zwei Semester in formalen Methoden.

Gezeigt werden sollte der Nutzen formaler Methoden im Lehrplan auf die erhöhte Fähigkeit der Studenten, komplexe Probleme zu lösen.

Die Studenten der CG wurden zufällig ausgewählt, die Mitglieder der FMG haben sich freiwillig gemeldet, sowohl für das zusätzliche Curriculum als auch für die Gruppe, die in dem Experiment die formalen Methoden anwendet.

Aufbau des Experiments:

1. Entwicklung einer Software für ein Fahrstuhlsystem
2. Abzugeben: Source Code und Executable
3. Optional: UML-Diagramm
4. FMG: Spezifikation
5. Gütekriterien:
   a) Executable erfüllt sechs Testfälle
   b) Knappheit des Codes
   c) Komplexität des Codes (Schleifen, Verschachtelung und Fallunterscheidungen)
6. Wissensstand und Fähigkeiten aller Studenten wurde durch vorherige Tests als gleich bezeichnet

Ergebnisse des Experiments:

1. Design der CG konnte nicht festgestellt werden, da keine UML-Diagramme abgegeben wurden
2. CG und FMG produzierten im Schnitt kaum optimalen Code, wobei der FMG-Code etwas besser war
3. Die FMG-Spezifikationen variierten in der Güte
4. Funktionale Korrektheit: 6/6 der FMG und 5/11 der CG

Im Anschluß führte ein 'VERIFICATION TEAM' (VT) aus vier Studenten der FMG eine vollständige formale Analyse (Spezifikation & Verifikation) durch und produzierten knappen, fehlerfreien Code und entdeckten eine Lücke in der verbalen Anforderungsbeschreibung.

Schlußfolgerungen aus dem Experiment:

1. Einsatz formaler Methoden produziert besseren Code
2. Einsatz formaler Methoden erhöht die Lösungsfähigkeit komplexer Probleme (complex problem solving skills)
3. Die Vermutung von Vertretern formaler Methoden (formale Methoden = bessere Software) ist bestätigt

Ein zweites Paper kritisiert das Experiment und die Schlußfolgerungen.

Vorwegnahme: Die Kritik ist berechtigt.

Welche Kritikpunkte gibt es?

Design des Experiments

CG und FMG sind ungleich: CG zufällig ausgewählt, FMG freiwillig
=> ungleiche Merkmale
=> Quasi-Experiment (Schlußfolgerungen sind nicht valide, solange es konkurrierende Hypothesen gibt)

FMG-Mitglieder eventuell höher motiviert

FMG-Mitglieder hatten mehr Lehrplan und Übungen
=> Mehr Praxis und Wissen

FMG-Mitglieder wurden zuvor einem 'Lernstil-Test' unterzogen und als kooperativ und ergebnisorientiert klassifiert.
Über die CG-Mitglieder liegt ein solcher Test nicht vor.

Dadurch konkurrierende Hypothesen:

a) FMG und CG sind gleich in Lernfähigkeit
b) FMG-Mitglieder sind bessere Lerner

FMG-Mitglieder sind eventuell begabter: Aus anderen Tests ging hervor, daß FMG-Mitglieder 30% bessere Leistungen erbrachten.

Hawthorne Effekt

Versuchspersonen, die wissen, daß sie beobachtet werden, verhalten sich anders als im „Normalzustand“, meist verhalten sie sich den (geahnten) Erwartungen der Versuchsleiter entsprechend.

Die Studenten wußten vermutlich, was das Ziel des Experiments war.

Mangelnde Kontrolle

Da die CG kein UML-Diagramm abgab, ist deren Arbeitsweise unbekannt.

Dadurch entstehen zwei konkurrierende Hypothesen:

a) CG hatte keinerlei Design genutzt sondern gleich codiert.
b) CG hatte formale, informelle, keine Analyse oder eine Mischung daraus genutzt.

Problem: Es liegen zu wenig Daten über die CG vor, so daß ein Vergleich zwischen CG und FMG nicht machbar ist.

Analyse der Resultate

Das Design eines Experiments ist abhängig von den aufgestellten Hypothesen.

In diesem Fall wurden Resultate genannt, die jeweils einen anderen Aufbau des Experiments verlangten.

Anlaß war, den Nutzen formaler Methoden im Lehrplan zu untersuchen. In den Schlußfolgerungen wurde allgemein der Nutzen formaler Methoden bestätigt.

Ähnliche Arbeiten

Auf Ergebnisse themengleicher Arbeiten wurde nicht eingegangen. So kam eine größere Studie auf andere Ergebnisse als im Experiment.

Zumindest die Abweichungen von anderen Ergebnissen hätte diskutiert werden sollen.

Externe Validierung

Durchführung und Resultate des Experiments hätte durch Dritte bewertet und diskutiert werden sollen.

Eventuelle Mängel des Experiments hätten angeführt werden müssen.

Statt dessen wurden Mängel vertuscht oder nebenbei abgehandelt bzw. in der Literaturliste versteckt.

Schlußfolgerung:

Ein Redesign ist zwingend notwendig. Zwar sind Fragestellung und Versuch interessant und wichtig, aber in der vorliegenden Form sind die Resultate und Schlußfolgerung unbrauchbar.

In einem Antwortpaper wiesen die Experimentatoren die Kritik mit markiger Sprache zurück. U.a. versuchten sie, den Hawthorne Effekt als widerlegt darzustellen, was aber nicht der Fall ist.

Wichtiger Grundsatz

Ehrlichkeit
+
korrekte Anwendung von Techniken
=
Vertrauenswürdigkeit

Fehler sind „erlaubt“, wichtig ist der aufrichtige Umgang mit Kritik.

Warum?

„It is impossible for every scientist to independently do every experiment to confirm every theory. Because life is short, scientists have to trust other scientists. So a scientist who claims to have done an experiment and obtained certain results will usually be believed, and most people will not bother to repeat the experiment.“

Richtlinien für „saubere“ Experimente

1. Zusammestellung von Versuchsgruppen: Zufällige Auswahl aus einer merkmals-gleichen Gruppe und zufällige Zuordnung zu Kontrollgruppe
   => Gleichverteilung der Merkmale in den Gruppen
2. Versuchsziel soll den Versuchspersonen unbekannt sein. Eventuell kann man durch Irreführung von den eigentlichen Zielen ablenken
   => Hawthorne Effekt verhindern
3. Vor Beginn Hypothesen aufstellen und nur daran Resultate herausziehen
   => Konsistenz im Experiment sicherstellen
4. Versuchsgruppen müssen in allen relevanten Bereichen überwacht werden
   => Vergleichbarkeit sichern

5. Themengleiche Arbeiten beachten
   => Auf erledigte Arbeit zurückgreifen
   => In bestehende Arbeiten einordnen
6. Experiment durch Dritte bestätigen bzw. überprüfen lassen
   => Finden von Fehlern
   => Akzeptanz in der scientific community
7. Aufbereitung des Experiments und ehrlich und ausführlich in den relevanten Bereichen
   => Akzeptanz in der scientific community
   /* Selbst das beste Experiment wird nicht angenommen und wird somit nutzlos, wenn es fehlerhaft oder unklar präsentiert wird */