Ein Prozeßmodell zur experimentellen Erprobung von Software-Entwicklungsprozessen

Oliver Laitenberger, Jürgen Münch

Ein Prozeßmodell zur experimentellen Erprobung von Software-Entwicklungsprozessen

Oliver Laitenberger[‡], Jürgen Münch[† *]
laiten@iese.fhg.de, muench@informatik.uni-kl.de

04/1996

Sonderforschungsbereich 501

[†]Arbeitsgruppe Software Engineering
[‡]Fraunhofer Institut für Experimentelles Software Engineering

Fachbereich Informatik
Universität Kaiserslautern
Postfach 3049
67653 Kaiserslautern
Germany

[*] In alphabetischer Reihenfolge.

Zusammenfassung

Mit Hilfe von Experimenten lassen sich systematisch Erkenntnisse über Software-Entwicklungsprozesse und Software-Entwicklungsumgebungen gewinnen, die als Folge eines Lernprozesses in zukünftige Software-Entwicklungsprojekte einfließen können. Ein grundlegendes Prinzip zur kontinuierlichen Verbesserung des Entwicklungsprozesses und der Entwicklungsumgebung ist das Quality Improvement Paradigm (QIP) [BCR94a]. Die Durchführung von Experimenten kann sowohl Bestandteil des QIP sein als auch selbst in die Vorgehensweise des QIP eingebettet werden. In diesem Dokument wird ein Prozeßmodell zur systematischen Durchführung von Experimenten im Software Engineering vorgestellt. Es basiert auf einer von Basili, Selby und Hutchens vorgeschlagenen Experimentiermethodik [BSH86], die verfeinert und in das QIP eingebettet wird. Dabei stehen die Anforderungen des Sonderforschungsbereichs 501 [SFB94] hinsichtlich der Durchführung von Experimenten im Vordergrund der Betrachtung. Der Schwerpunkt dieses Dokuments liegt in der Darstellung der bei der Durchführung von Experimenten notwendigen Schritte und Aktivitäten. Bezüglich der Ausführung der Aktivitäten innerhalb eines Schrittes kann es bei der konkreten Durchführung von Experimenten Wahlmöglichkeiten geben.

Inhaltsverzeichnis

1. Einleitung

1.1 Experimentelles Software Engineering

Die Forschung in der Informatik beschäftigt sich in vielen Fällen mit der Entwicklung von neuen Systemen, Techniken oder Werkzeugen. Dabei wird oft nicht explizit gemacht, aus welchem Kontext das System, die Technik oder das Werkzeug stammt, welcher Nutzen in verschiedenen Kontexten erbracht wird bzw. welche Eigenschaften das System, die Technik oder das Werkzeug besitzt, die es besser als andere Ansätze erscheinen läßt. Eine Möglichkeit der systematischen Gewinnung von Erkenntnissen über verschiedener Systeme, Techniken oder Werkzeuge besteht in der Durchführung von Experimenten. In anderen Wissenschaften wie z. B. der Medizin oder der Soziologie gehören Experimente zum wissenschaftlichen Standardrepertoire. Im Vergleich dazu sind die analytischen Verfahren in der Informatik noch nicht ausreichend ausgeprägt. Es mangelt an Vorgehensweisen, wie Experimente zu planen, durchzuführen und auszuwerten sind. Dieses Dokument faßt die bisher als wesentlich erkannten Schritte des Experimentierens im Bereich Software Engineering zusammen und paßt sie an die Bedürfnisse des Sonderforschungsbereichs 501 (im folgenden kurz SFB genannt) [SFB94] an. Ausgangspunkt ist dabei eine Vorgehensweise, die von Basili, Selby und Hutchens [BSH86] vorgeschlagen worden ist.

Ein auf experimentell gewonnenen Erfahrungen basierender Einsatz von Systemen, Techniken oder Werkzeugen ist in Software-Entwicklungsorganisationen bisher nicht weit verbreitet [Gla94]. Vielfach werden Techniken angewandt, ohne ihre Eignung für einen bestimmten Entwicklungskontext (Erfahrung der Entwickler, Anwendungsdomäne etc.) zu kennen bzw. zu berücksichtigen. Entsprechendes Wissen wird nicht ermittelt und in weiteren Projekten verwendet. Die meisten Organisationen lernen nicht systematisch aus durchgeführten Projekten, um die gewonnenen Erkenntnisse in nachfolgenden Projekten sinnvoll umzusetzen. So gehen wertvolle Erfahrungen verloren, die zur Verbesserung des Entwicklungsprozesses und somit zur Wettbewerbsfähigkeit der Organisationen beitragen könnten.

Wesentliche Gründe für die mangelnde Bereitschaft zur Durchführung von Experimenten in der Praxis sowie die unzureichende Fähigkeit, aus durchgeführten Projekten zu lernen, sind zum einen die hohen Kosten, die mit Experimenten verbunden sind, und zum anderen ein Zeitmangel in realen Projekten, der Lernaktivitäten nicht zuläßt. Doch hat sich bereits in einigen Software-Entwicklungsorganisationen (z. B.: NASA [NAS94], CSC [BM95]) die Erkenntnis durchgesetzt, daß Investitionen hierfür besser sind als das Eingehen von Risiken beim Gebrauch neuer Werkzeuge oder Techniken, von denen man keinerlei Kenntnis über die mit ihrem Einsatz verbundenen Konsequenzen hat. Um der Bedeutung des Lernvorgangs gerecht zu werden, sollte er in einer organisatorischen Struktur innerhalb einer Software-Entwicklungsorganisation institutionalisiert sein. Ein Ansatz hierzu ist die „Experience Factory“ [BCR94a], die die Analyse und Wiederverwendung von in Projekten gesammelten Erfahrungen unterstützt und von der Projektorganisation, die das Software-System erstellt, organisatorisch getrennt ist.

Es gibt verschiedene Software-Entwicklungsprozesse, mit denen qualitativ hochwertige Software-Systeme entwickelt werden können. Da es weder einen universalen Entwicklungsprozeß noch eine optimale Entwicklungsmethode gibt, muß der Entwicklungsprozeß immer in dem ihn umgebenden Kontext betrachtet werden. Die nichtrationale Auswahl von Techniken, Methoden und Werkzeugen, möglicherweise intuitiv, führt selten zu einer optimalen Lösung. Wichtig ist die systematische Auswahl und Anpas-

sung von Software-Entwicklungsprozessen im Hinblick auf die von der Entwicklungsorganisation angestrebten Ziele. Dabei muß beachtet werden, daß der Entwicklungsprozeß von Projekt zu Projekt eine veränderliche Größe ist. Lernen, wie Technologien an die Bedürfnisse von Software-Entwicklungsorganisationen angepaßt werden, umfaßt die Durchführung von Experimenten verschiedener Typen (Fallstudien, replizierte Projekte etc.).

1.2 Experimenteller Ansatz des SFB 501

Bei der Entwicklung großer Systeme innerhalb des SFB sollen Softwareprodukte, Entwicklungsschritte und Erfahrungen aus abgelaufenen Projekten mit Hilfe von generischen Methoden[1] wiederverwendet werden. Dies setzt voraus, daß während der Durchführung von Software-Entwicklungsprojekten schritthaltend Informationen gesammelt werden. Solche Informationen sind u. a. Teilprodukte, Entwicklungsschritte oder Ergebnisse von Messungen mit den zugehörigen Begründungen, warum sie in dem jeweiligen Projekt verwendet bzw. durchgeführt wurden. In ersten Projekten werden solche Begründungen auf bereits vorhandenen Erfahrungen, zur Verfügung stehenden Informationen oder subjektiven Einschätzungen beruhen, in späteren Projekten auf dem dokumentierten Wissen und Erfahrungen aus vorausgegangenen Projekten.

Um den Einsatz von bestimmten Technologien innerhalb des SFB mittelfristig objektiv begründen zu können und um langfristig eine generische Methodik für ihre Auswahl und Anpassung an eine gegebene Situation zu entwickeln, werden Software-Entwicklungsprojekte innerhalb des SFB grundsätzlich als Experimente verstanden. Die Software-Entwicklung wird im Kontext des SFB als Labordisziplin verstanden. Entsprechend umfaßt die Rahmenarchitektur des SFB neben Projektorganisationen, die Software-Systeme erstellen, einen Software-Engineering-Kern (kurz: SE-Kern), in dem Informationen und Erfahrungen (in Form von Produktmodellen, Prozeßmodellen, Vorgehensmodellen, Qualitätsmodellen und Beschreibungstechniken) aus vergangenen Projekten analysiert, aufbereitet und gespeichert werden und in einem Rückkopplungsprozeß zur Verbesserung der Modelle und Beschreibungstechniken zukünftigen Projekten zur Verfügung gestellt werden.

Der SE-Kern spiegelt sich im Software-Engineering-Labor (kurz: SE-Labor, Teilprojekt A1) des SFB wider. Hier sammeln sich methodische Ergebnisse und sonstige Erkenntnisse des SFB an. Bereits vorhandene bzw. neu entwickelte Werkzeuge oder Techniken müssen erprobt werden, um sie in das Labor zu integrieren. Das SE-Labor institutionalisiert somit den projektübergreifenden Lernprozeß innerhalb des SFB. Die Entwicklung der prototypischen Anwendungssysteme in den Projektorganisationen erfolgt unter Benutzung des SE-Labors. Zusammen mit Teilbereich D (Prototypanwendungen) bildet das SE-Labor das experimentelle Umfeld für Entwicklungsprojekte des SFB.

Im Projektbereich B des SFB werden die methodischen Grundlagen für die Entwicklung großer Systeme erarbeitet. Hierzu gehört insbesondere die Erarbeitung von aufeinander abgestimmten Produkt-, Prozeß- und Qualitätsmodellen, die Entwicklung von generischen Methoden zur Software-Entwicklung und deren Einbettung in ein übergeordnetes Rahmenmodell.

1. Unter dem Begriff „generische Methoden“ werden im SFB alle Beschreibungs- und Generierungstechniken mit den dazugehörigen Werkzeugen zusammengefaßt, durch die eine Wiederverwendung von existierenden Softwareprodukten, Entwicklungsschritten und Erfahrungen bei der Entwicklung eines neuen Systems methodisch unterstützt wird.

Der Projektbereich C beschäftigt sich mit Beschreibungstechniken, die zur Erfassung aller relevanten Aspekte der im Projektbereich B zu entwickelnden Modelle und Methoden geeignet sind.

1.3 Ziele

Die Entwicklung von Modellen, Methoden und Techniken in den Projektbereichen B und C sollte einerseits auf bereits gemachten Erfahrungen beruhen, andererseits auch durch den Einsatz in zukünftigen Projekten später beeinflußt werden. Idealerweise sollte ein Modell, eine Methode oder eine Technik nicht zur Wiederverwendung aufbereitet werden, bevor sie in einem realistischen Projektkontext ausprobiert wurde. Diese Erprobung kann mit Hilfe von Experimenten erfolgen, durch die sich systematisch Erkenntnisse gewinnen lassen. Dies erfordert ein Prozeßmodell zur Durchführung von Experimenten im Software Engineering im allgemeinen, welches für den SFB instantiiert werden kann. Ein solches Prozeßmodell wird in diesem Dokument vorgestellt. Es ist als initiales Modell zu verstehen, das mit Hilfe von Erfahrungen in durchgeführten Experimenten verfeinert bzw. modifiziert werden soll. Mit ihm werden folgende Ziele verfolgt:

- *Vorgabe einer einheitlichen Vorgehensweise zur systematischen Durchführung von Experimenten im Software Engineering.* Durch die Einheitlichkeit des Vorgehens bezüglich bestimmter Schritte auf einer abstrakten Ebene wird ein Vergleich von Experimenten sowie ihre Wiederholung vereinfacht. Erfahrungen mit der Vorgehensweise sowie daraus resultierende Verbesserungen kommen allen Experimenten zugute. Die Systematik hat einen wesentlichen Einfluß auf den Wert der bei den Experimenten erzielten Resultate.
- *Beitrag zur methodischen und theoretischen Fundierung des SE-Kerns*. Der Lernprozeß innerhalb des SFB bzw. allgemein innerhalb von Software-Entwicklungsorganisationen ist iterativ. Somit gehört das Sammeln von Erfahrungen über Modelle, Methoden bzw. Beschreibungstechniken zu einem festen Bestandteil eines kontinuierlichen Verbesserungsprozesses. Das hier vorgestellte Prozeßmodell kann später in ein umfassendes Rahmenmodell zur Modellierung aller die Wiederverwendung unterstützender Aspekte der Software-Entwicklung integriert werden.
- *Verfeinerung und Formalisierung der Schritte des Quality Improvement Paradigm (QIP)*. Das wesentliche Prinzip, das den experimentellen Ansatz des SFB begründet, ist die kontinuierliche Verbesserung des Entwicklungsprozesses und der Entwicklungsumgebung. Eine Formulierung dieses Prinzips ist das QIP [BCR94a], das in diesem Dokument zugrunde gelegt wird. Die Anwendung des QIP bei der Durchführung von Projekten, d. h. auf der konstruktiven Seite der Software-Entwicklung, ist bereits vielfach untersucht worden. Der Einsatz des QIP im Rahmen von Experimenten, d. h. auf der analytischen Seite der Software-Entwicklung, fand bisher jedoch kaum Beachtung. Insbesondere die kontinuierliche Verbesserung einer Experimentiermethodik schritthaltend mit durchgeführten Experimenten sowie das systematische Sammeln von Erfahrungen beim Experimentieren ist wenig untersucht. Ein Ziel dieses Dokuments ist es, hierzu einen Beitrag zu leisten, um zu einer Verfeinerung und Formalisierung der Schritte des QIP zu gelangen.

Die Vorstellung des Prozeßmodells zur Durchführung von Experimenten legt einen Schwerpunkt in der Darstellung der notwendigen Schritte und Aktivitäten. Eine detaillierte Beschreibung der einzelnen Aktivitäten erfolgt in der referenzierten Literatur.

Das Dokument wurde im Teilprojekt B1 im Rahmen des Arbeitspaketes 6 (Formulierung einer Methode zur experimentellen Erprobung von Software-Entwicklungsprozessen) erstellt und wird im Rahmen des Arbeitspaketes 11 (Schritthaltende Integration der Techniken und Werkzeuge im SE-Labor) in das SE-Labor integriert.

1.4 Aufbau des Dokuments

Das Dokument ist wie folgt aufgebaut: In Kapitel 2 werden allgemeine Eigenschaften von Experimenten erläutert. Kapitel 3 ordnet die Vorgehensweise beim Experimentieren den einzelnen Schritten des QIP zu. Die Beschreibung des Prozeßmodells erfolgt zunächst auf abstraktem Niveau (Kapitel 4) und wird danach entsprechend den Schritten des QIP verfeinert (Kapitel 5-10). Zu jedem Schritt werden die Vorbedingungen, das Vorgehen, die Nachbedingungen, beteiligte Rollen, ein Beispiel und Literaturreferenzen angegeben. Als durchgängiges Beispiel wird ein Experiment zum Vergleich von Testtechniken verwendet [BS87], wobei eine Vereinfachung und Anpassung einzelner Teile an den hier vorgegebenen Rahmen vorgenommen wurde. Wesentliche Begriffe dieses Dokuments werden im Anhang A erläutert.

1.5 Benutzung des Dokuments

Dieses Dokument stellt einen Überblick für diejenigen dar, die sich bisher noch nicht oder wenig mit Experimenten im Software Engineering befaßt haben. Es liefert darüber hinaus eine Anleitung zur Durchführung von Experimenten im SFB, wobei ein Rahmenmodell vorgestellt wird, innerhalb dessen verschiedene Wahlmöglichkeiten bezüglich der Durchführung von konkreten Experimenten bestehen.

Die Durchführung von Experimenten umfaßt eine Vielzahl von Aktivitäten, die von verschiedenen Personen durchgeführt werden können. Durch die Zusammenfassung von zusammengehörenden Aktivitäten kann man Rollen identifizieren. Je nach Rolle gibt es verschiedene Sichten bei der Durchführung von Experimenten, die mit verschiedenen Informationen, Kenntnissen und Interessen verbunden sind. Für dieses Dokument werden folgende Rollen im Zusammenhang mit der Durchführung von Experimenten als sinnvoll erachtet:

- *Experimentator.* Die Aufgaben des Experimentators umfassen die Planung eines konkreten Experiments (im Rahmen des hier vorgestellten Prozeßmodells), die Koordination der einzelnen Aktivitäten bei der Durchführung des Experiments sowie die Auswertung des Experiments.
- *Qualitätsmanager.* Der Qualitätsmanager hat die Aufgabe, Wissen und Erfahrungen (in Form von Modellen, Methoden, Beschreibungstechniken etc.) zu verwalten, sie in konkreten Experimenten während der Planung und Durchführung zur Verfügung zu stellen und sie aufgrund von Resultaten aus Experimenten zu verbessern.
- *Experimentteilnehmer.* Der Untersuchungsgegenstand eines Experiments kann sich auf eine Aktivität bzw. eine Menge von Aktivitäten sowie ein Produkt bzw. eine Menge von Produkten beziehen. Experimentteilnehmer sind bei der Durchführung dieser Aktivitäten bzw. bei

der Erstellung dieser Produkte beteiligt. Ihr Interesse muß nicht notwendigerweise auf die Ziele des Experiments ausgerichtet sein, sondern es kann z. B. ausschließlich ein zu erstellendes Software-System betreffen.

Bei der Beschreibung der einzelnen Schritte wird jeweils im Unterpunkt „Beteiligte Rollen“ erläutert, welche Rollen am entsprechenden Schritt beteiligt sind. Dementsprechend können Schwerpunkte bei der Arbeit mit diesem Dokument gelegt werden. Es ist jedoch sicherlich sinnvoll, daß alle Rollen einen Überblick über die Vorgehensweise haben (dieser Überblick wird in den Kapiteln 3 und 4 vermittelt).

Folgende Zuordnung der Rollen innerhalb des SFB wird als sinnvoll erachtet: Der Experimentator kommt aus den Teilbereichen B oder C mit der Intention, ein Methode, Technik etc. in einem experimentellen Umfeld zu analysieren. Die Rolle des Qualitätsmanagers wird vom SE-Labor wahrgenommen. Experimentteilnehmer können Mitglieder von Projektteams des SFB oder Mitglieder von speziell für ein Experiment zusammengestellten Gruppen sein.

Auch andere Zuordnungen sind möglich: Möchte z. B. das SE-Labor die Eignung von Werkzeugen für bestimmte Zwecke in konkreten Umgebungen ermitteln, so sollte es zusätzlich die Rolle des Experimentators wahrnehmen. Die tatsächliche Rollenverteilung ist zwischen den an einem Experiment beteiligten Personen zu vereinbaren. Hierbei kann es vorkommen, daß eine Rolle von mehreren Personen ausgeübt wird, umgekehrt kann eine Person auch mehrere Rollen ausüben.

2. Allgemeine Eigenschaften von Experimenten

2.1 Einleitung

In diesem Kapitel werden wichtige Eigenschaften von Experimenten diskutiert, die bei der Planung (insbesondere dem Entwurf) neuer Experimente berücksichtigt werden sollen, da sie einen Einfluß auf die Akzeptanz der experimentellen Vorgehensweise sowie auf den Wert der experimentellen Ergebnisse haben.

2.2 Skalierbarkeit

Die Skalierbarkeit ermöglicht es, Experimente zunächst „in kleinem Rahmen“ durchzuführen, um so erste Ergebnisse zu bekommen und zu gewährleisten, daß mit dem Experiment tatsächlich die Hypothese überprüft werden kann. Vorteil davon ist, daß das Experiment im kleinen Rahmen billiger ist und sich einfacher durchführen läßt. Beispielsweise kann ein Experiment erst im universitären Umfeld durchgeführt werden, um erste Ergebnisse zur experimentellen Hypothese und auch zum Experiment selbst zu erhalten, bevor das Experiment im industriellen Rahmen ausgeführt wird. Ein Experiment sollte keine einmalige Sache darstellen. Vielmehr sollte ein Experiment so geplant werden, daß es einfach erweiterbar (skalierbar) ist. Skalierbarkeit bezeichnet die Eigenschaft, daß der Umfang (engl.: scope; siehe Tabelle 1) eines Experiments erweitert werden kann.

Für experimentelle Studien im Bereich der Informatik sind dabei zwei Aspekte besonders wichtig [BSH86]:

1. Anzahl der Teams, die untersucht werden;
2. Anzahl der Projekte, an denen die Teams arbeiten.

„Teams“ sind Gruppen, die unabhängig voneinander arbeiten. Ein Team kann auch aus nur einer einzigen Person bestehen. Teams haben bestimmte Eigenschaften wie z. B. Erfahrung oder Organisationsform. „Projekte“ sind Entwicklungsprogramme oder Problemstellungen, an denen die Teams arbeiten. Projekte haben ebenfalls Eigenschaften wie z. B. Größe oder Anwendungsbereich. Die Eigenschaften der Teams und Projekte müssen beim Experimentieren berücksichtigt werden.

Die Aufteilung von experimentellen Studien nach Teams und Projekten ermöglicht die folgende Klassifikation des Umfangs von Experimenten:

		Anzahl der Projekte	
		1	> 1
Anzahl der Teams	1	Single Project	Multi-Project Variation
	> 1	Replicated Project	Blocked Subject-Project

Tabelle 1: Umfang (engl.: scope) [BSH86]

„Blocked Subject-Project“-Studien untersuchen mehrere Teams, die an mehreren Projekten arbeiten. „Replicated Project“-Studien untersuchen Teams, die an einem Projekt arbeiten, während „Multi-Project Variation“-Studien ein Team bei der Bearbeitung von mehreren Projekten untersuchen. „Single-Project“-Studien untersuchen ein Team, das an einem Projekt arbeitet.

In der Literatur gibt es keine einheitliche Begriffsbildung für den Typ eines Experiments. Pfleeger unterscheidet z. B. „Fallstudien“ und „formale Experimente“ in Abhängigkeit der ausgeübten Kontrolle [Pfl94]. Hiernach fallen „Single Projekt“-Studien in die Kategorie einer Fallstudie und „Blocked Subject-Project“-Studien in die Kategorie eines formalen Experiments.

Die Qualität der Ergebnisse, aber auch die Kosten des Experiments, steigen von links nach rechts und von oben nach unten (siehe Tabelle 1). Skalierbarkeit ermöglicht es, zuerst eine Fallstudie durchzuführen und diese dann entsprechend zu erweitern. Dadurch wird das Risiko reduziert, in einem aufwendigen Experiment ungültige Ergebnisse zu erhalten.

2.3 Wiederholbarkeit

Neben der Skalierbarkeit spielt die Wiederholbarkeit eine wichtige Rolle. Wiederholbarkeit bezieht sich dabei auf die Tatsache, daß die Bedingungen, unter denen ein Experiment ausgeführt worden ist, bei einer Wiederholung ebenfalls hergestellt werden können. Es gibt verschieden Gründe für Wiederholbarkeit:

- Durch Wiederholung können Fehler, die beim Experimentieren gemacht worden sind, besser eingeschätzt werden.
- Wiederholung ermöglicht es, die einzelnen Einflußfaktoren besser beurteilen zu können.
- Das Vertrauen in die Ergebnisse kann erhöht werden.

In vielen Fällen bekommt man bei einmaliger Durchführung von Experimenten nicht genügend Datenpunkte, um statistisch signifikante Ergebnisse abzuleiten. Sofern die Bedingungen konstant gehalten werden, kann durch Wiederholung die Anzahl der Datenpunkte erhöht werden, um somit statistisch signifikante Ergebnisse abzuleiten.

Um Wiederholbarkeit zu ermöglichen, müssen die Bedingungen, unter denen ein Experiment ausgeführt wird, genau beschrieben werden. Die genaue Beschreibung ermöglicht es auch Personen, die nicht am Experiment teilgenommen oder es geplant haben, dieses in ihrer Umgebung durchzuführen und die Ergebnisse zu vergleichen.

Eine Wiederholung kann Nebenprodukt eines Folgeexperiments sein, bei dem eine neue Fragestellung untersucht wird. Allerdings müssen dieselbe Vorgehensweise verwendet und dieselben Bedingungen garantiert werden wie beim ursprünglichen Experiment.

2.4 Kosten

Die Durchführung eines Experiments ist in den meisten Fällen mit hohen Kosten verbunden. Dies sind nicht nur die Kosten, die durch den Zeitaufwand der Teilnehmer anfallen, sondern auch die Zeit, die in Planung, Entwurf und Analyse des Experiments investiert werden müssen. Dies ist unter anderem ein Grund, warum bisher so wenige Experimente durchgeführt worden sind. Neben der Unkenntnis, wie überhaupt Experimente zu gestalten sind, werden oft die Kosten gescheut.

Da der Kostenaspekt sehr wichtig ist, sollten Daten gesammelt werden, die verläßliche Aussagen über die Kosten des Experiments ermöglichen. Dadurch, daß verläßliche Informationen vorhanden sind, wieviel Aufwand ein Experiment tatsächlich verursacht hat, kann besser beurteilt werden, ob ein Experiment wiederholt werden kann (insbesondere von Personen, die nicht am Experiment beteiligt waren).

2.5 Kontrolle

Das Ziel von Experimenten ist es, einen Zusammenhang zwischen verschiedenen Faktoren zu untersuchen. Dabei unterscheidet man zwischen Faktoren, die beobachtet bzw. manipuliert werden (unabhängige Variablen), um ihren Einfluß auf diejenigen Faktoren, die gemessen werden (abhängige Variablen), zu überprüfen. Unabhängige Variablen, die manipuliert werden können, bezeichnet man auch als kontrollierte Variablen. Die Ergebnisse eines Experiments werden unbrauchbar, wenn sich erst nach Durchführung des Experiments herausstellt, daß sich die Werte der abhängigen Variablen nicht allein durch die Werte der unabhängigen Variablen erklären lassen, sondern auch nicht-erfaßte Variablen die abhängigen Variablen beeinflußt haben. Deshalb müssen alle Variablen, die einen signifikanten Einfluß auf die Ergebnisse haben könnten, beobachtet werden, um gültige Aussagen zu machen.

Fallstudien lassen sich von formalen Experimenten dadurch unterscheiden, daß der Grad der Kontrolle von abhängigen Variablen bei Fallstudien sehr viel geringer (bzw. gar nicht vorhanden) ist als bei formalen Experimenten [Pfl94]. Fallstudien können als ein spezieller Experimenttyp betrachtet werden. Aufgrund der Problematik bei der Kontrolle lassen sich die Ergebnisse von Fallstudien schwieriger verallgemeinern als die von kontrollierten Experimenten (auch hierbei ist Vorsicht angeraten). Eine Fallstudie kann ein erster Schritt sein, um ein kontrolliertes Experiment vorzubereiten (siehe Skalierbarkeit). Darüber hinaus kann eine Fallstudie in Rahmen eines realistischen Projekts auch zur Validierung von Resultaten dienen, die in einem experimentellen Umfeld ermittelt wurden.

Vollständige Kontrolle aller unabhängigen Variablen ist in den meisten Fällen schwierig zu erreichen und muß für jedes Experiment getrennt diskutiert werden. Möglichkeiten zur Minimierung von Einflüssen bestehen z. B. darin, daß entweder Teams gebildet werden, die möglichst homogen sind („Blocking“) oder die Experimentteilnehmer den Teams zufällig zugewiesen werden („Randomization“).

2.6 Validität

Im Zusammenhang mit Experimenten spielt die Validität eine entscheidende Rolle für die Akzeptanz der Ergebnisse. Dabei läßt sich zwischen drei Arten von Validität unterscheiden [JSK91]:

1. Validität des Experiments (engl.: construct validity)
   Bei dieser Art von Validität wird danach gefragt, wie die theoretischen Überlegungen, die meist in einer Hypothese formalisiert werden, im Experiment tatsächlich überprüft und gemessen werden können.
2. Interne Validität (engl.: internal validity)
   Bei dieser Art der Validität wird danach gefragt, im welchem Umfang es das Experiment ermöglicht, einen kausalen Zusammenhang zwischen den abhängigen und unabhängigen Variablen zu messen.
3. Externe Validität (engl.: external validity)
   Bei dieser Art der Validität wird danach gefragt, wie sich die im Experiment gewonnen Ergebnisse verallgemeinern lassen.

Beim Experimentieren ist insbesondere darauf zu achten, daß die interne Validität gewährleistet ist. Es gibt folgende Gefahren für die interne Validität (siehe hierzu [JSK91]):

1. Auswahl (engl.: Selection)
   Es ist darauf zu achten, daß sich die Experimentteilnehmer nur in bezug auf die unabhängigen Variablen signifikant unterscheiden. Ansonsten kann es vorkommen, daß Unterschiede im Experiment auf Faktoren zurückzuführen sind, die nicht beachtet und auch nicht gemessen wurden. Die Ergebnisse sind dann weniger gültig.

2. Veränderungen der Experimentteilnehmer (engl: Maturation)
   Experimentteilnehmer können sich während eines Experiments ändern. In diesem Zusammenhang sind Lerneffekte zu betrachten. Lerneffekte beziehen sich auf die Tatsache, daß Experimentteilnehmer während des Experiments hinzulernen und sich die Ergebnisse des Experiments nicht aufgrund der Variation einer unabhängigen Variablen verändern, sondern aufgrund des Lerneffekts.
   Unter den Aspekt „Veränderung der Experimentteilnehmer“ fällt auch, daß Experimentteilnehmer während des Experiments ihr Verhalten ändern können. Verhaltensänderungen treten z. B. durch Langeweile, Ermüdung oder auch durch die einfache Beobachtung der Experimentteilnehmer (Hawthorne-Effekt [Par74]) auf.

3. Geschichte (engl.: History)
   Jedes Ereignis, das im Zeitraum der Experimentdurchführung stattfindet, kann die Ergebnisse beeinflussen. Dies kann z. B. die Krankheit eines Experimentteilnehmers, die Ersetzung eines Experimentteilnehmers durch einen anderen oder eine Verzögerung durch einen Systemabsturz sein.

4. Instrumentierung (engl.: Instrumentation)
   Jede Änderung in der Art und Weise, wie die Daten des Experiments erfaßt werden, kann einen Einfluß auf die Ergebnisse haben. Deshalb sollten die Art und Weise der Datenerfassung im Verlauf des Experiments nicht verändert werden. Ein Beispiel hierfür wäre die Veränderung eines Fragebogens während der Ausführung des Experiments. Oftmals kommt es zu Änderungen der Instrumentierung erst bei Wiederholungen des Experiments, da entsprechende Details des Ausgangsexperiments nicht bekannt sind. In solchen Fällen muß man mit dem Vergleich von Ergebnissen vorsichtig sein. Aufgrund der Bedeutung von Wiederholungen für die Gültigkeit der Ergebnisse sollte auf eine einheitliche Instrumentierung geachtet werden.

3. Einordnung der Durchführung von Experimenten in das Quality Improvement Paradigm (QIP)

Das Quality Improvement Paradigm (QIP) [Bas95][BCR94a], ursprünglich entwickelt als Ansatz zur systematischen Verbesserung von Software-Projekten, kann auch auf die Durchführung von Experimenten angewendet werden. Die folgende Tabelle stellt dar, wie die Aktivitäten der einzelnen Schritte des QIP bei Projekten beziehungsweise bei der Durchführung von Experimenten instantiiert werden.

Schritt	Projekt	Experiment
1	Charakterisieren des Projekts und der entsprechenden Projektumgebung	Charakterisieren des Experiments und der Umgebung, in der das Experiment stattfinden soll
2	Setzen von Projektzielen	Aufstellen von Zielen und Ableiten von entsprechenden Hypothesen
3	Auswählen einer geeigneten Vorgehensweise; erstellen des Projektplans	Wählen bzw. Entwickeln des Entwurfs für das Experiment
4	Ausführen des Projekts	Durchführen des Experiments
5	Analyse der Projektergebnisse	Analyse der Daten; Annahme oder Ablehnung der Hypothese; Ableiten von neuen oder modifizierten Hypothesen
6	Sicherung der Erfahrungen	1. Sicherung von Erfahrungen über das Experiment als solches 2. Sicherung von Ergebnissen des Experiments

Tabelle 2: Vergleich der Schritte des QIP bei der Durchführung von Projekten und Experimenten

Das QIP von Projekten wird in [Bas95] näher erläutert, so daß hier nicht mehr explizit darauf eingegangen werden soll. Der erste Schritt der Anwendung des QIP auf Experimente besteht darin, den experimentellen Rahmen zu beschreiben, in dem das Experiment stattfinden soll. Dies beinhaltet die Charakterisierung sowohl des Experiments als auch der Umgebung, in die das Experiment eingebettet ist. Zur Charakterisierung gehört die Identifikation von Schwachstellen, die verbessert werden können. Aus der Charakterisierung im ersten Schritt werden im Schritt 2 die Ziele des Experiments extrahiert. Die Ziele legen fest, was mit dem Experiment erreicht werden soll und warum es ausgeführt wird. Nachdem die Ziele festgelegt sind, werden die Hypothesen für das Experiment aufgestellt. Die Hypothesen bestimmen, was beim Experiment untersucht werden soll. Nachdem Ziel und Hypothesen fixiert sind, wird im Schritt 3 des QIP der Entwurf für das Experiment ausgewählt bzw. entwickelt. Der Entwurf soll eine Überprüfung der Hypothesen ermöglichen. Schritt 4 besteht aus der Durchführung des Experiments gemäß des in Schritt 3 spezifizierten Entwurfs. Während der Durchführung werden die Daten gesammelt, mit denen eine Überprüfung der Hypothesen möglich ist. Die gesammelten Daten werden dann im Schritt 5 analysiert und statistisch ausgewertet. Die Anwendung von statistischen Tests führt hierbei zur Annahme bzw. Ablehnung der Hypothesen. Im letzten Schritt (Schritt 6) werden die gesammelten Erfahrungen gesichert. Dabei sollten sowohl die Ergebnisse des Experiments als auch die Erfahrungen über das Experimentieren als solches dokumentiert werden. In den meisten Fällen wird man bei der Durchführung eines Experiments auf neue, offene Fragestellungen oder Hypothesen aufmerksam. Diese sollten ebenfalls in Schritt 6 dokumentiert werden.

Die oben beschriebenen Schritte, die durch das QIP definiert sind, beschreiben keine starre sequentielle Abfolge der Aktivitäten. Vielmehr gibt es Abhängigkeiten zwischen den Aktivitäten, die zu Zyklen bei der Anwendung führen können. So legt z. B. der Entwurf die Klasse von statistischen Testverfahren fest, mit denen die Hypothesen überprüft werden können. Will man einen bestimmten Test bei der Analyse anwenden, so erzwingt dies eine bestimmte Art von Entwurf. Eine Aktivität muß nicht unbedingt erst nach Beendigung einer anderen ausgeführt werden, sondern es kann zu Überlappungen von Aktivitäten kommen. So kann z. B. das Sichern von Erfahrungen schritthaltend mit der Durchführung des Experiments erfolgen. Allerdings sollten vor der Durchführung eines Experiments (Schritt 4) die Schritte 2 und 3 zwingend abgeschlossen sein. Bei der Ausführung einer Aktivität kann es auch zu Lerneffekten kommen, die eine erneute Bearbeitung eines vorhergehenden Schrittes notwendig macht. Deshalb ist es beim Experimentieren wichtig, einen Überblick über alle wesentlichen Aktivitäten zu haben. Das in diesem Dokument vorgestellte Prozeßmodell identifiziert und beschreibt dazu alle Aktivitäten, die im Verlauf eines Experiments auszuführen sind und verfeinert somit die einzelnen Schritte des QIP.

Experimente können in Projekte eingebettet sein. Dabei werden innerhalb eines umfassenden Projekts ein oder mehrere Experimente abgewickelt, so daß es zu einer Verzahnung der Vorgehensweisen für die Durchführung von Projekten und Experimenten gemäß dem QIP kommen kann (siehe Abbildung 1). Auch die umgekehrte Verzahnung der Vorgehensweisen ist möglich, falls verschiedene Projekte in ein Experiment einbezogen sind.

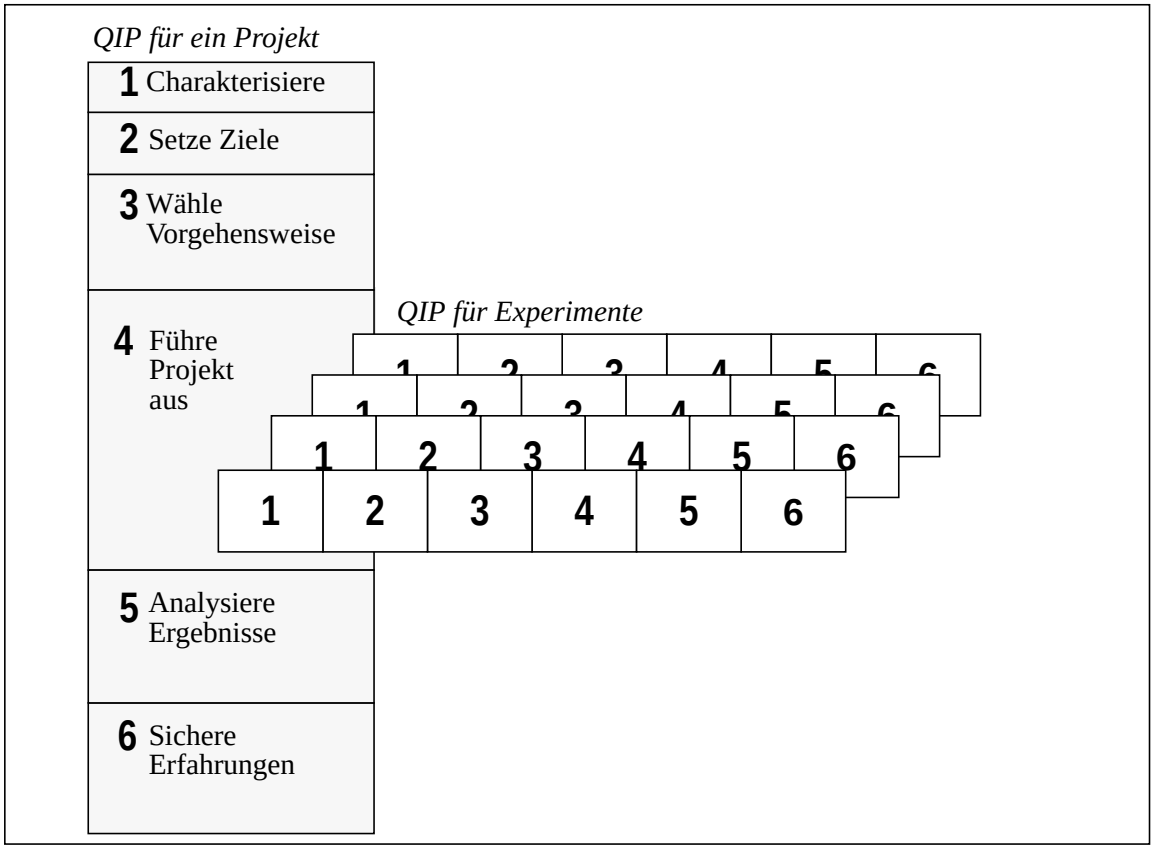

Abbildung 1: Verzahnung von Experimenten und Projekten

4. Top-Level-Sicht des Prozeßmodells

Die einheitliche Beschreibung von Experimenten, die durch das Prozeßmodell festgelegt ist, erleichtert die Wiederholung von Experimenten und den Vergleich zwischen Experimenten. Dabei können nicht nur die Ergebnisse des gesamten Experiments, sondern auch die Ergebnisse jedes Schrittes und den damit verbundenen Aktivitäten verglichen werden.

In diesem Kapitel wird eine Übersicht des Prozeßmodells auf dem obersten Abstraktionsniveau in Form einer grafischen Darstellung gegeben (siehe Abbildung 2[1]). Dabei werden Prozesse, Produkte und Beziehungen zwischen ihnen dargestellt. Der Begriff „Prozeß“ wird hier im Sinne einer Aktivität verwendet, der Begriff „Produkt“ im Sinne eines Artefakts (z. B.: Informelle Beschreibung, Entwurfsdokument), das bei der Durchführung von Experimenten anfällt bzw. benötigt wird. Man spricht auch von konsumierten und produzierten bzw. modifizierten Produkten.

Eine detaillierte Beschreibung der einzelnen Prozesse und Produkte findet sich in den folgenden Kapiteln. Die verfeinerten Prozesse und Produkte aus Abbildung 2 (Schritt 2, Schritt 3, GQM-Plan, Entwurfsdokument) werden in den entsprechenden Kapiteln mit einem höheren Verfeinerungsgrad grafisch dargestellt.

1. Die Form der graphischen Darstellung orientiert sich an [BD+95].

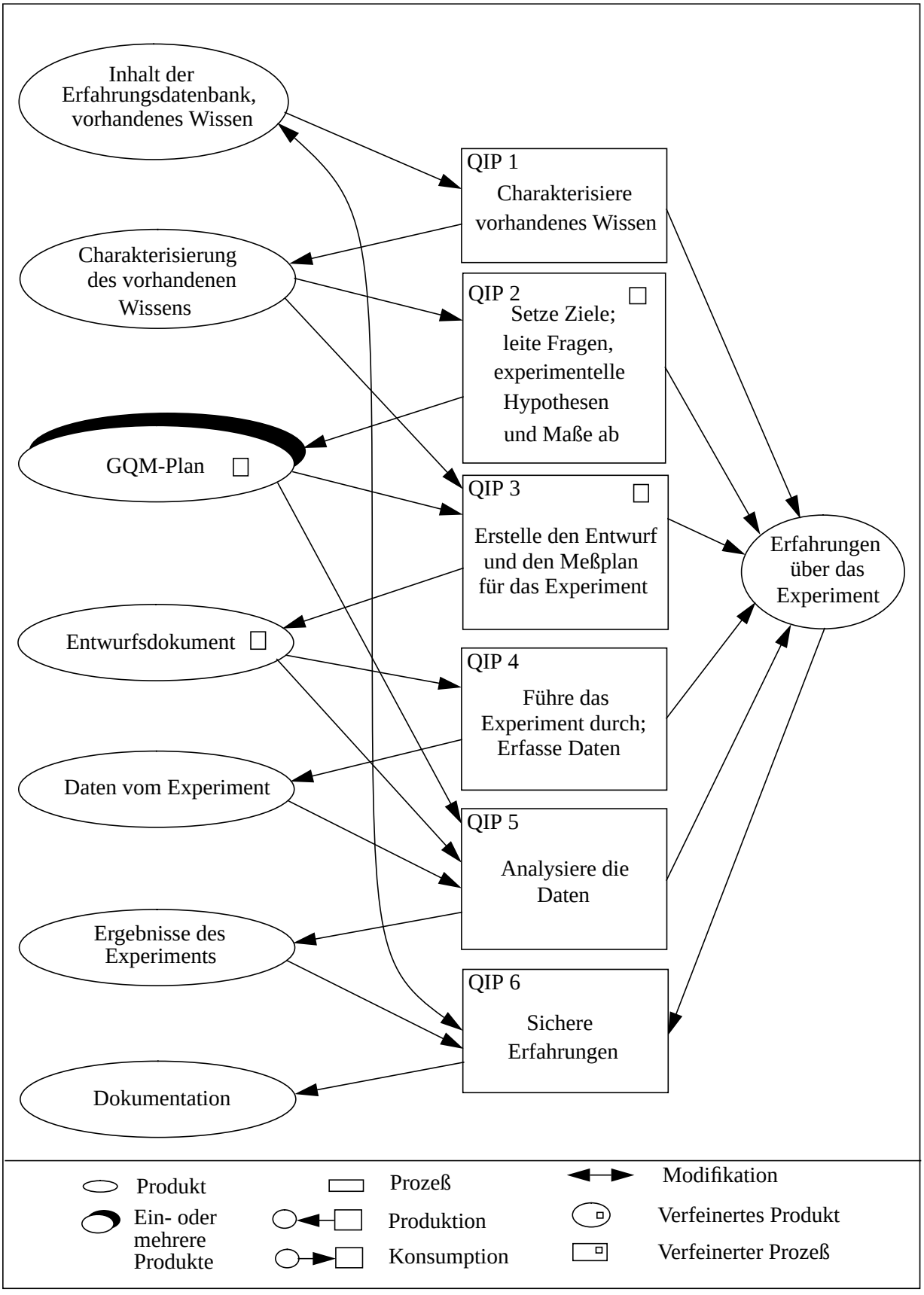

Abbildung 2: Prozeßmodell für Experimente (Top-Level)

5. Charakterisierung vorhandenen Wissens (QIP 1)

5.1 Vorbedingungen

Die Vorbedingung besteht in der Verfügbarkeit des Zugriffs auf die für diesen Schritt notwendigen Informationsquellen. Dabei reichen diese Informationsquellen von Literaturbeschaffung und -recherche bis zum Suchen von entsprechenden Einträgen in einer Erfahrungsdatenbank (sofern vorhanden) [BCR94a].

5.2 Vorgehen

Die folgende Tabelle faßt die wichtigsten Aspekte dieses QIP-Schrittes zusammen:

QIP1: Charakterisiere vorhandenes Wissen			
Wissen über das Objekt		Wissen über das Experiment	
Erfahrungsdatenbank	Objektfaktoren	Motivation	Experimentfaktoren
Prozeßmodelle	Eigenschaften	Verstehen	Probleme
Produktmodelle	Vermutungen	Bewerten	Ressourcen
Qualitätsmodelle	Fragestellungen	Vergleichen	Prozeß
GQM-Pläne		Verbessern	Validität
		Lernen	Relevante Objektfaktoren

Tabelle 3: Schritt 1 des QIP: Charakterisierung vorhandenen Wissens

Der erste Schritt des QIP besteht in der Charakterisierung des vorhandenen Wissens. Diese Charakterisierung dient als Grundlage zum Aufstellen von Zielen, zum systematischen Ableiten von experimentellen Hypothesen und zum Erstellen des experimentellen Entwurfs in den weiteren QIP-Schritten. Die im folgenden angegebenen Aspekte geben einen Überblick der inhaltlichen Bestandteile des vorhandenen Wissens. Über den Umfang der Charakterisierung kann keine allgemeingültige Aussage gemacht werden, da er von den mit dem Experiment verfolgten Zwecken abhängt.

Bei der Charakterisierung des vorhandenen Wissens wird zwischen dem bereits vorhandenen Wissen über das Objekt, das untersucht werden soll, und dem bereits vorhandenen Wissen über das Experiment, das durchgeführt werden soll, unterschieden. Das Wort Objekt soll hier als Synonym für die zu untersuchende Technik, Methode, Werkzeug, Modell, Prozeß etc. verwendet werden. Innerhalb des Experiments kann es ein oder mehrere Objekte geben.

Vorhandenes Wissen über das Objekt wird aus der Erfahrungsdatenbank extrahiert. Wenn keine Erfahrungsdatenbank existiert, kann dieses Wissen auch über andere Informationsquellen erschlossen werden. Es umfaßt bereits vorhandene Prozeßmodelle, Qualitätsmodelle, Produktmodelle und GQM-Pläne[1]. Objektfaktoren sind Fragestellungen, Vermutungen über das Objekt oder Eigenschaften des Objekts. Dies kann auch Schwachstellen des Objekts beinhalten. Dabei ist nicht relevant, ob die Objektfaktoren in direktem Zusammenhang mit dem geplanten Experiment stehen oder nicht. Es kann sich z. B. um Erfahrungen handeln, die bei der Entwicklung einer Technik, Methode oder eines Werkzeugs gemacht wurden, im

1. GQM-Pläne werden im zweiten QIP-Schritt (siehe Kapitel 6) zur systematischen Ableitung experimenteller Hypothesen und Maße verwendet und sind dort näher erklärt.

geplanten Experiment allerdings nicht überprüft werden sollen. Auch das vorhandene Wissen über unterschiedliche Eigenschaften des Objekts in verschiedenen Entwicklungskontexten wird hier beschrieben[1].

Die Charakterisierung des Wissens über das Experiment besteht zum einen aus der Motivation für das Experiment, zum anderen aus den Experimentfaktoren.

Der Punkt Motivation soll beschreiben, warum das Experiment ausgeführt wird. Mögliche Motivationen sind, einen Vergleich zwischen verschiedenen Objekten durchzuführen, mehr über ein Objekt zu lernen, etc.

Als Experimentfaktoren werden hier alle Aspekte bezeichnet, die das konkrete Experiment und seine Umgebung charakterisieren. Dies umfaßt die Faktoren Probleme, Ressourcen, Prozeß, Qualität und relevante Objektfaktoren.

Der Faktor Probleme stellt Einschränkungen dar, aus denen Probleme entstehen können, wenn sie nicht berücksichtigt werden. Einschränkungen können sich z. B. aus dem Anwendungsgebiet ergeben, wenn eine Technik nur für Echtzeitsysteme verwendet werden kann, die Teilnehmer sich allerdings mit dieser Art von Systemen nicht auskennen.

Der Faktor Ressourcen läßt sich unterteilen in Teilnehmer, Ort, Zeitrahmen, Budget und Werkzeuge. Unter Teilnehmer[2] werden alle Aspekte beschrieben, die im Zusammenhang mit den Teilnehmern des Experiments stehen. Unter Ort wird beschrieben, wo das Experiment ausgeführt werden soll. Der Zeitrahmen legt sowohl die Dauer des Experimentes insgesamt als auch die Zeit, in der die Teilnehmer benötigt werden, fest. Teilnehmer und Zeit bestimmen im wesentlichen das Budget, das ebenfalls unter dem Punkt Ressourcen beschrieben wird. Der letzte Punkt des Faktors Ressourcen besteht in der Beschreibung der zur Verfügung stehenden Werkzeuge (z. B.: statistische Werkzeuge).

Das experimentelle Prozeßmodell und Gültigkeitsaspekte werden durch die Faktoren Prozeß und Validität beschrieben.

Der Faktor Prozeß beschreibt das gewählte Prozeßmodell, anhand dessen das Experiment durchgeführt wird, und, soweit bereits im voraus getroffen, Entscheidungen bezüglich einer konkreten Vorgehensweise. Dieser Bericht enthält ein Prozeßmodell, innerhalb dessen noch Wahlmöglichkeiten bezüglich einzelner Schritte bestehen. Eine Wahlmöglichkeit besteht z. B. bei der Auswahl bzw. Erstellung des Entwurfs des Experiments.

Unter dem Faktor „Validität“ werden Aussagen bezüglich der Validität zusammengefaßt. Hier kann z. B. beschrieben werden, welche externe Validität das Experiment haben soll.

Unter den relevanten Objektfaktoren werden alle Eigenschaften, Vermutungen und Fragestellungen des Objekts verstanden, die das Experiment beeinflussen bzw. die durch das Experiment untersucht werden sollen. Die relevanten Objektfaktoren sind eine Untermenge der Objektfaktoren. Während bei den Objekt-

1. Der Grund für die umfassende Beschreibung der Objektfaktoren ist die Charakterisierung des gesamten Wissens über das Objekt (möglicherweise in verschiedenen Entwicklungskontexten). Die Beschreibung kann darüber hinaus als Ausgangspunkt für die Durchführung weiterer Experimente dienen. Die Durchführung von weiteren Experimenten trägt dann iterativ zur Wissensgewinnung über das Objekt bei.
2. Die Teilnehmer eines Experiments werden auch als Subjekte des Experiments bezeichnet.

faktoren alle Eigenschaften, Vermutungen und Fragestellungen zu dokumentieren sind, beschreiben die relevanten Objektfaktoren lediglich diejenigen Eigenschaften, Vermutungen und Fragestellungen, die für das Experiment relevant sind.

5.3 Nachbedingungen

Die Nachbedingung für den ersten Schritt ist die Existenz einer Charakterisierung des vorhandenen Wissens bezüglich des Objekts (bzw. der Objekte) und des Experiments. Diese Beschreibung sollte vollständig sein in dem Sinne, daß alle in Tabelle 3 angegebenen Punkte darin berücksichtigt werden.

5.4 Beteiligte Rollen

Bei der Ausführung dieses Schrittes sind der Experimentator und Qualitätsmanager beteiligt. Der Qualitätsmanager hilft dem Experimentator, das Wissen über das Objekt und das Wissen über das Experiment aus der Erfahrungsdatenbank zu extrahieren[1]. Sofern Experten bezüglich des Objektes zur Verfügung stehen, können auch von ihnen Auskünfte eingeholt werden. Der Experimentator wählt die für ihn nützlichen Informationen aus und erstellt ein Dokument, in dem das vorhandene Wissen beschrieben ist.

5.5 Beispiel

Als Beispiel für die Vorgehensweise beim Experimentieren soll hier ein Experiment verwendet werden, das die Effektivität von Testtechniken überprüft. Dieses Experiment wurde 1982 von Basili/Selby [BS87] durchgeführt. Anhand dieses Beispiels wird die Vorgehensweise durchgängig erläutert.

Die Objekte, die untersucht werden sollten, waren drei Testtechniken: Strukturelles Testen, funktionales Testen und Code reading. Zum Zeitpunkt des Experiments existierte noch keine Erfahrungsdatenbank, so daß das Wissen über die Objekte aus der über Testtechniken vorhandenen Literatur extrahiert werden mußte. Dabei wurde vor allem das Wissen über die Aktivitäten, die bei jeder Testtechnik auszuführen sind, und bereits vorhanden Qualitätsmodelle, die den Aufwand der Anwendung einer Testtechnik beschreiben, wiederverwendet. Objektfaktoren in diesem Beispiel sind der Aufwand zum Finden und Isolieren von Fehlern, der Aufwand zum Erlernen einer Testtechnik, die Effizienz der Testtechniken, die Fehlerklassen, die mit einer Testtechnik gefunden werden können, etc. Eine Vermutung, die in den Objektfaktoren dokumentiert sein könnte, wäre: „Der Aufwand zum Erlernen von Code reading ist geringer als zum Erlernen von strukturellem oder funktionalem Testen“.

Die Motivation für das Experiment war die Durchführung eines Vergleichs zwischen den drei Testtechniken. Das Experiment wurde zum einen mit Studenten der Universität Maryland, zum anderen mit professionellen Entwicklern der NASA durchgeführt. Dabei mußten mehrere Probleme gelöst werden: Erstens sind die Kenntnisse von Studenten über Testtechniken sehr unterschiedlich, so daß sich bereits dadurch unterschiedliche Ergebnisse ergeben. Zweitens mußten Beispielprogramme verwendet werden, die sowohl beim Experiment mit den Studenten als auch beim Experiment mit den professionellen Entwicklern, die normalerweise Software für Satelliten entwickeln, verwendet werden können. Der Aspekt Teilnehmer des Faktors Ressource wurde bereits beschrieben (Studenten und professionelle Entwickler der NASA). Das

1. Falls noch keine Erfahrungsdatenbank zur Verfügung steht, beschränkt sich die Wissensbeschaffung auf andere Quellen (Literatur, Berichte, erzählte Erfahrungen etc.). Hierbei kann der Qualitätsmanager unterstützend tätig sein.

Experiment fand an der Universität Maryland und in Räumlichkeiten der NASA statt. Das Budget für das Experiment wurde nicht explizit beschrieben. Zur Auswertung der Daten waren diverse statistische Werkzeuge vorhanden. Als Zeitrahmen für die Durchführung standen mehrere Tage zur Verfügung. Als Prozeßmodell wurde eine Vorgehensweise verwendet, die in [BSH86] [Sel93] beschrieben ist. Der Entwurf des Experiments (Fractional-Factorial Design [BHH78]) wurde ursprünglich in den Sozialwissenschaften angewendet und garantierte zusammen mit der Analysetechnik (Analysis of Variance) die interne Validität der Ergebnisse des Experiments. Im Experiment selbst wurden der Aufwand zum Finden und Isolieren von Fehlern, die Effizienz der Testtechniken und die Fehlerklassen, die mit einer Testtechnik gefunden werden können, untersucht. Diese Aspekte werden unter „relevante Objektfaktoren“ aufgeführt.

5.6 Literatur

[BCR94a]
Victor R. Basili, Gianluigi Caldiera, and H. Dieter Rombach
The Experience Factory
Encyclopedia of Software Engineering (John J. Marciniak, ed.), vol. 1, pp. 469-476, John Wiley & Sons, 1994.

[BHH78]
G. E. P. Box, W. G. Hunter, and J. S. Hunter
Statistics for Experimenters
John Wiley & Sons, New York, 1978.

[BS87]
Victor R. Basili and Richard W. Selby
Comparing the Effectiveness of Software Testing Techniques
IEEE Transactions on Software Engineering, vol. 13(12), pp. 1278-1296, December 1987.

[BSH86]
Victor R. Basili, Richard W. Selby, and David H. Hutchens
Experimentation in Software Engineering
IEEE Transactions on Software Engineering, vol. SE-12, pp. 733-743, July 1986.

[Sel93]
Richard W. Selby
Software Measurement and Experimentation Frameworks, Mechanisms, and Infrastructure
in Experimental Software Engineering Issues: A critical assessment and future directions, H. D. Rombach, V. R. Basili, and R. W. Selby, eds., pp. 89-106, Lecture Notes in Computer Science Nr. 706, Springer-Verlag, September 1993.

[Sha88]
Richard J. Shavelson
Statistical Reasoning for the Behavioral Sciences
2. ed., 1988, Allyn & Bacon, Inc., ISBN 0-205-11765-1.

6. Aufstellen von Zielen und Ableiten von Hypothesen (QIP 2)

6.1 Vorbedingungen

Die Vorbedingung besteht in der Existenz einer möglichst genauen Charakterisierung des experimentellen Rahmens. Diese Charakterisierung soll alle wichtigen Informationen enthalten, die im ersten QIP-Schritt des Prozeßmodells aufgeführt sind. Sie wird im zweiten QIP-Schritt durch Anwendung des GQM-Ansatzes [Bas92] formalisiert und verfeinert, so daß experimentelle Hypothesen und Maße abgeleitet werden können.

6.2 Vorgehen

Die folgende Tabelle faßt die wichtigsten Aspekte dieses QIP-Schrittes zusammen:

QIP2: Aufstellen von Zielen und Ableiten von Hypothesen		
Experimentelles Ziel	GQM-Plan	Hypothese(n)
Zweck	Ziele	experimentelle Hypothese(n)
Objekt	Fragen	Nullhypothese(n)
Aspekt	Maße	Alternativhypothese(n)
Blickwinkel	Skalen	
Kontext		

Tabelle 4: Schritt 2 des QIP: Aufstellen von Zielen und Ableiten von experimentellen Hypothesen

Die erste Aktivität besteht in der Definition der experimentellen Ziele. Hierunter werden diejenigen Ziele verstanden, an denen der Experimentator interessiert ist. Unabhängig davon können die Experimentteilnehmer und der Qualitätsmanager andere Ziele verfolgen, die mit dem Experiment nicht unmittelbar zusammenhängen müssen (z. B. können Experimentteilnehmer nur an der Erstellung von Produkten interessiert sein). Die nächste Aktivität besteht darin, aus den experimentellen Zielen die experimentellen Hypothesen abzuleiten, die im Experiment untersucht werden sollen. Als letzte Aktivität dieses QIP-Schritts werden Maße bestimmt, deren Ausprägungen die Entscheidung ermöglichen, ob eine Hypothese angenommen oder abgelehnt wird. Diese Ausprägungen werden bei der Durchführung von Messungen im Rahmen des 4. QIP-Schritts ermittelt.

Für die zielgerichtete Bestimmung von geeigneten Maßen existieren verschiedene Top-down-Strategien. In [Rom90] werden drei solcher Verfahren, Quality Function Deployment (QFD), Software Quality Metrics (SQM) und Goal/Question/Metric (GQM), miteinander verglichen. Dabei wird festgestellt, daß der QFD-Ansatz auf die Identifizierung der Charakteristika von Software-Produkten zielt, der SQM-Ansatz für die Zertifizierung bestimmter Qualitäten des auszuliefernden Endproduktes gedacht ist und der GQM-Ansatz es ermöglicht, jedes beliebige Meßziel zu unterstützen. Beim GQM-Ansatz gibt es „keine Begrenzungen hinsichtlich des zu messenden Objekts, des Meßzwecks, des Aspekts, der gemessen werden soll, und des Blickwinkels von Interesse“ [Rom90]. Da der GQM-Ansatz der flexibelste Ansatz hinsichtlich der möglichen Analyseziele ist, wird er in diesem Dokument verwendet.

Der Goal/Question/Metric-Ansatz (GQM-Ansatz)

Der Goal/Question/Metric-Ansatz [Bas92][BCR94b] unterstützt die zielgerichtete Auswahl von Maßen, indem er eine systematische Vorgehensweise anbietet, um Meßziele zu definieren und auf operationale Weise zu verfeinern. Mit ihm lassen sich Ziele umfassend definieren, indem durch eine explizite Top-down-Verfeinerung Fragen abgeleitet und daraus Maße bestimmt werden.

Bei der Anwendung des GQM-Ansatzes im Rahmen dieses Dokuments werden experimentelle Hypothesen mit einbezogen. Hierbei ist zu erwähnen, daß im Rahmen der Weiterentwicklung des GQM-Ansatzes an der Universität Kaiserslautern der Begriff der Hypothese konzeptionell bereits belegt ist [Mar95] und teilweise mit dem hier verwendeten Konzept der experimentellen Hypothese nicht übereinstimmt[1]. Innerhalb dieses Dokuments ist die Verwendung des Begriffs Hypothese immer im Sinne einer experimentellen Hypothese zu verstehen.

Im folgenden werden die einzelnen Aktivitäten des 2. QIP-Schrittes beschrieben, wobei der GQM-Ansatz unter Einbeziehung von experimentellen Hypothesen zugrunde gelegt wird.

Definition von Zielen

Für die Definition des Ziels steht im GQM-Ansatz ein Template zur Verfügung [Rom90]. Dieses Template besteht aus fünf Punkten:

Analysiere

{*Objekte*: den Komponententest, den Entwurfsprozeß, das Werkzeug, ...}

zum Zwecke des

{*Warum*: Charakterisierens, Verstehens, Bewertens, Vergleichs, ...}

bezüglich

{*Aspekte*: der Kosten, der Korrektheit, der Änderbarkeit, der Zuverlässigkeit, ...}

aus dem Blickwinkel

{*Wer*: des Entwicklers, des Testers, des Managers, ...}

im Kontext

{*Organisation*: Universität S, Firma T, ...

Personen: professionelle Entwickler, umgeschulte Programmierer, ...

Methoden und Werkzeuge: Testtechnik U, Debugger V, Beschreibungsmittel W, ...

Umfang: „Single Project“-Experiment, „Replicated-Project“-Experiment,

„Multi-Project-Variation“-Experiment“, „Blocked Subject-Project“-Experiment

1. Unterschiede zwischen einer experimentellen Hypothese und einer Hypothese des GQM-Ansatzes, wie sie in [Mar95] definiert wurde, bestehen vor allem in der Definition, der Motivation und der Verwendung von Hypothesen innerhalb des GQM-Ansatzes. Eine Integration der beiden Konzepte ist Gegenstand künftiger Forschungsaktivitäten.

weitere Kontextfaktoren: Anwendungsbereich X, Entwicklungsprozeß Y, ...}

Die für dieses Template notwendigen Informationen sollten aus der Charakterisierung des vorhandenen Wissens abgeleitet werden, die im ersten QIP-Schritt des Prozeßmodells erstellt worden ist. Dabei sind die *Objekte* und möglicherweise auch die *Aspekte* bereits durch das Objektwissens charakterisiert, der *Zweck* kann aus der Motivation abgeleitet werden. Die Festlegung des *Blickwinkels* auf das zu untersuchende Objekt sollte vom Experimentator vorgenommen werden. Hierbei kann die Charakterisierung des Objektwissens Anregungen geben, sofern sie Informationen über die Bedeutung verschiedener Eigenschaften des Objekts bezüglich unterschiedlicher Blickwinkel enthält. Der *Kontext* ist in den Experimentfaktoren beschrieben, wobei der *Umfang* in Abhängigkeit von der gewünschten Validität des Experiments sowie vorgegebenen Ressourcenbeschränkungen (z. B.: Budgetgrenzen, verfügbares Personal) festzulegen ist.

Ableiten von Fragen und experimentellen Hypothesen

Aus dem im ersten Schritt definierten experimentellen Ziel werden anhand von Richtlinien [Rom90] Fragen abgeleitet. Die Schwierigkeit besteht darin, eine möglichst sinnvolle Abdeckung der Ziele durch Fragen zu erreichen. Aus den bisherigen Erfahrungen bezüglich der Anwendung des GQM-Ansatzes wurde ersichtlich, daß es hilfreich ist, die Fragen in GQM-Plänen hierarchisch zu strukturieren. Dies bedeutet, daß verschiedene Fragen unter einer übergeordneten Frage gruppiert werden können [Mar95].

Da die Fragen über die Maße in direktem Zusammenhang mit den erhobenen Daten stehen, sollte bei der Durchführung dieses QIP-Schrittes sorgfältig gearbeitet werden. Stellt sich nach dem Experiment heraus, daß für die Erfüllung des Zieles zusätzliche Fragen und damit auch zusätzliche Maße notwendig sind, ist die Gültigkeit der Resultate zumindest teilweise in Frage zu stellen.

Den Fragen können experimentelle Hypothesen zugeordnet werden. Eine experimentelle Hypothese besteht aus einer Vermutung, die man über das Objekt (bzw. die Objekte) hat. Es ist auch möglich, daß sich eine experimentelle Hypothese auf mehrere Fragen bezieht.

Für die Formulierung der experimentellen Hypothesen gibt es folgende Richtlinien:

1. Eine experimentelle Hypothese sollte eine bestimmte Beziehung zwischen verschiedenen Eigenschaften von Objekten (bzw. eines Objekts) ausdrücken.
2. Eine experimentelle Hypothese sollte einfach und eindeutig formuliert sein.
3. Eine experimentelle Hypothese sollte testbar sein, d. h., daß die Hypothese anhand von gesammelten Daten überprüfbar sein sollte.

Hat man z. B. die Frage: „Können mit Testtechnik A mehr Fehler gefunden werden als mit Testtechnik B?“, so läßt sich daraus die folgende experimentelle Hypothese ableiten: „Mit Testtechnik A werden mehr Fehler gefunden als mit Testtechnik B“.

Aus einer experimentellen Hypothese werden die Nullhypothese und die Alternativhypothese abgeleitet. Der Grund dafür ist, daß in der Statistik zwischen Nullhypothese und Alternativhypothese unterschieden und mit der Nullhypothese gearbeitet wird. Der charakteristische Unterschied besteht darin, daß die Nullhypothese solange als wahr angenommen wird, bis die erfaßten Daten dagegen sprechen. Somit konzentriert sich ein Experiment eher auf Abweichungen von der Nullhypothese als auf Abweichungen von

der Alternativhypothese. In diesem Sinne bedeutet „Testen der experimentellen Hypothese“, zu bestimmen, ob die erfaßten Daten überzeugend genug sind, um die Nullhypothese zu verwerfen und die Alternativhypothese als wahr zu akzeptieren [Pfl95a].

Ausgangspunkt ist die experimentelle Hypothese. Die Nullhypothese ist die Negation der experimentellen Hypothese. Die Nullhypothese für das obige Beispiel lautet also: „Mit Technik A werden weniger oder genau so viele Fehler gefunden wie mit Technik B“. Die Alternativhypothese stimmt mit der experimentelle Hypothese überein. Durch die Daten, die im Experiment gesammelt werden, soll statistisch gezeigt werden, daß die Wahrscheinlichkeit, daß die Nullhypothese tatsächlich zutrifft, sehr gering ist. Ist diese Wahrscheinlichkeit kleiner als eine festgelegte Grenze, die als Signifikanzniveau bezeichnet wird, so läßt sich mit hoher Wahrscheinlichkeit die Nullhypothese ablehnen und somit die Alternativhypothese, die der ursprünglichen Vermutung entspricht, annehmen. Für eine genauere Diskussion des Zusammenhanges zwischen experimenteller Hypothese, Nullhypothese und Alternativhypothese wird auf die Literatur verwiesen (z. B.: [JSK91] [Sha88]).

Bestimmung von Maßen

Der letzte Aktivität besteht in der Bestimmung von Maßen, mit denen die Fragen beantwortet und die experimentellen Hypothesen überprüft werden können. Bei der Top-down-Verfeinerung eines Ziels werden in einem letzten Schritt aus einer detaillierten Frage diejenigen Maße abgeleitet, die es erlauben, die Frage zu beantworten. Bezüglich der Verfeinerung von Fragen gibt es die Restriktion, daß eine Frage nicht gleichzeitig durch Maße und Fragen auf einer Ebene verfeinert werden darf. Diese Restriktion wurde eingeführt, damit es für alle Maße eine eindeutige Motivation gibt [Mar95]. Es ist möglich, daß gleiche Maße verschiedenen Fragen beantworten.

Bei der Bestimmung der Maße ist auf die Festlegung der richtigen Skala zu achten. Insbesondere dürfen später lediglich die Operationen auf Maßen ausgeführt werden, die für die Skala auch erlaubt sind.

Es werden die folgenden Skalen mit ihren jeweiligen Operationen unterschieden [Cur80] [Fen91] [Fen94]:

Skala	Erlaubte Operationen	Beschreibung	Beispiel
Nominal	=, ≠	Kategorien	Geschlecht, Rasse
Ordinal	=, ≠, <, >	Rangordnungen	Schulnoten
Intervall	=, ≠, <, >, +, -	Gleiche Intervalle	Temperatur (^{0}F, ^{0}C),
Rational	=, ≠, <, >, +, -, /, *	Gleiche Intervalle und absolute Null	Gewicht in Kilogramm, Temperatur (K)

Tabelle 5: Verschiedene Skalen

GQM-Plan

Die Beziehungen zwischen einem Ziel, Fragen (mit evtl. zugeordneten experimentellen Hypothesen) und Maßen können als gerichteter, azyklischer Graph dargestellt werden, der dann einen *GQM-Plan* darstellt. Abbildung 3 zeigt die Struktur eines GQM-Plans. Es wird ein Ziel gezeigt, das durch zwei Fragen verfeinert ist. Die Fragen werden durch Maße verfeinert. Dabei trägt z. B. Maß 2.1 zur Beantwortung der Fragen 1.1 und 2.1 bei.

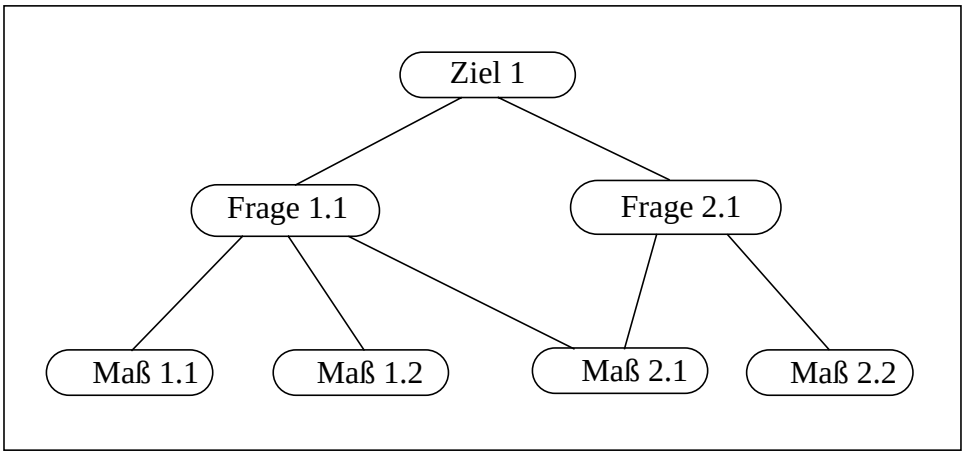

Abbildung 3: Hierarchischer Aufbau eines GQM-Plans

Die folgende Abbildung faßt die Aktivitäten und Produkte des zweiten QIP-Schritts des Prozeßmodells noch einmal zusammen:

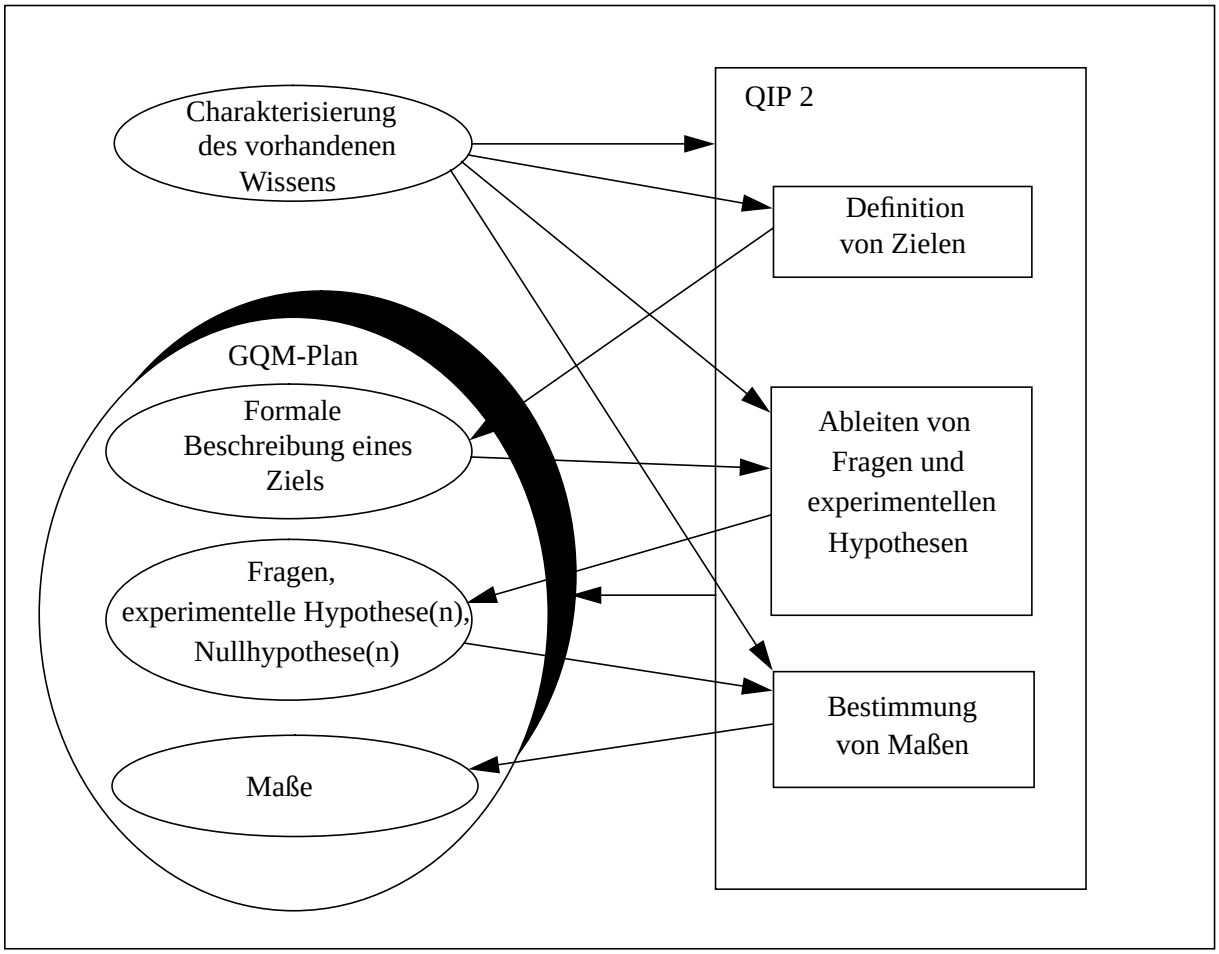

Abbildung 4: Verfeinerung des Schrittes 2 des QIP

6.3 Nachbedingungen

Nach der kompletten Ausführung der Aktivitäten dieses Schrittes sollte pro Ziel ein detaillierter GQM-Plan zur Verfügung stehen. Es ist möglich, daß dieser Schritt iteriert wird. Dies ist insbesondere dann der Fall, wenn die GQM-Pläne nach einer Iteration nicht vollständig sind, da möglicherweise noch Aspekte späterer Schritte berücksichtigt werden müssen (z. B. wenn im Entwurf festgestellt wird, daß ein Maß vergessen wurde oder weitere Hypothesen untersucht werden sollen).

6.4 Beteiligte Rollen

Für die Durchführung dieses Schrittes ist der Experimentator zuständig. Er führt alle hier aufgeführten Aktivitäten durch, wobei Experten, die sich mit dem Objekt bzw. der Anwendungsdomäne auskennen, sowie Experten, die Erfahrungen im Umgang mit dem GQM-Ansatz haben, hinzugezogen werden sollten.

6.5 Beispiel

Für das Testexperiment von Basili und Selby [BS87] läßt sich das Ziel wie folgt formulieren:

Analysiere

die Testtechniken funktionales Testen, strukturelles Testen und Code reading (*Objekte*)

zum Zwecke des

Vergleichs (*Warum*)

bezüglich

der Testeffektivität (*Aspekte*)

aus dem Blickwinkel

des Entwicklers (*Wer*)

im Kontext

der Universität Maryland und der NASA (*Organisation*)

mit Studenten und NASA-Entwicklern (*Personen*)

in einem „Blocked Subject-Project“-Experiment (*Umfang*).

Mögliche *Fragen* zur Quantifizierung dieses Zieles sind[1]:

1. Mit welcher der Testtechniken wurde die größte Anzahl von Fehlern gefunden?
2. Welche der Testtechniken hat die höchste Fehlererkennungsrate (Anzahl der gefundenen Fehler dividiert durch die dafür benötigte Zeit)?

Den Fragen lassen sich (möglichst basierend auf vorhandenen Vermutungen) z. B. die folgenden *experimentellen Hypothesen* zuordnen:

1. Die hier aufgeführten Aspekte des Ziels sowie die Fragen sind nicht vollständig und dienen nur beispielhaft der Veranschaulichung der Vorgehensweise. Auch ist die Anzahl der hier vorgestellten experimentellen Hypothesen nur eine Teilmenge der in diesem Experiment tatsächlich untersuchten experimentellen Hypothesen.

1. Mit strukturellem Testen werden mehr Fehler gefunden als mit funktionalem Testen.
2. Die Fehlererkennungsrate ist bei strukturellem und funktionalem Testen gleich oder bei ersterem geringer.

Daraus lassen sich dann die *Nullhypothesen* ableiten (die *Alternativhypothesen* entsprechen den experimentellen Hypothesen):

1. Nullhypothese H_0: Mit strukturellem Testen werden weniger oder gleich viele Fehler gefunden als mit funktionalem Testen.
   Alternativhypothese H_1: Mit strukturellem Testen werden mehr Fehler gefunden als mit funktionalem Testen.
2. Nullhypothese H_0': Die Fehlererkennungsrate ist bei strukturellem Testen höher als bei funktionalem Testen.
   Alternativhypothese H_1': Die Fehlererkennungsrate ist bei strukturellem und funktionalem Testen gleich oder bei ersterem geringer.

Der letzte Schritt besteht in der Bestimmung der *Maße*. Die Daten, die zur Beantwortung der ersten Frage und somit zur Überprüfung der ersten experimentellen Hypothese gesammelt werden müssen, sind die Anzahl der Fehler, die von jedem Teilnehmer in jedem Projekt gefunden werden. Die Daten, die zur Beantwortung der zweiten Frage und somit zur Überprüfung der zweiten experimentellen Hypothese gesammelt werden müssen, sind die Anzahl der Fehler und die Zeit, die die Teilnehmer benötigen, um die Fehler zu finden. Beide Maße sind rational-skaliert, weshalb im zweiten Fall die Division der Anzahl an Fehlern durch die Zeit erlaubt ist.

6.6 Literatur

[Bas92]
Victor R. Basili
Software Modeling and Measurement: The Goal/Question/Metric Paradigm
Technical Report CS-TR-2956, Department of Computer Science, University of Maryland, College Park, MD, 20742, September 1992.

[BCR94b]
Victor R. Basili, Gianluigi Caldiera, and H. Dieter Rombach,
The Goal Question Metric Approach
Encyclopedia of Software Engineering (John J. Marciniak, ed.), vol. 1, pp. 528-532,
John Wiley & Sons, 1994.

[BS87]
Victor R. Basili and Richard W. Selby
Comparing the Effectiveness of Software Testing Techniques
IEEE Transactions on Software Engineering, vol. 13(12), pp. 1278-1296, December 1987.

[Cur80]
Bill Curtis
Measurement and Experimentation in Software Engineering
Proceedings of the IEEE, vol. 68, pp. 1144-1157, September 1980.

[Hoi94]
Barbara Hoisl
A Process Model for Planning GQM-based Measurement
Technical Report STTI-94-06-E, Software-Technology-Transfer-Initiative Kaiserslautern, University of Kaiserslautern, 67653 Kaiserslautern, Germany, April 1994.

[Fen91]
Norman E. Fenton
Software Metrics: A Rigorous Approach
Chapman & Hall, London, 1991.

[Fen94]
Norman E. Fenton
Software Measurement: A Necessary Scientific Basis
IEEE Transactions on Software Engineering, vol. 20, nr. 3, pp. 199-206, March 1994.

[JSK91]
Charles Judd, Eliot R. Smith, and Louise Kidder
Research Methods in Social Relations
6th ed., Harcourt Brace Jovanovich College Publishers, ISBN 0-03-031149-7.

[Mar95]
Marco van Maris
GQM-DIVA: Ein Werkzeug zur Definition, Interpretation und Validation von GQM-Plänen
Diplomarbeit, Universität Kaiserslautern, Fachbereich Informatik, Mai 1995.

[Pfl95a]
Shari L. Pfleeger
Experimental Design and Analysis in Software Engineering
Part 2: How to Set Up an Experiment
ACM SIGSOFT Software Engineering Notes, vol. 20, no. 1, January 1995.

[Rom90]
H. Dieter Rombach
Practical Benefits of Goal-oriented Measurement
Proceedings of the Annual Workshop of the Centre for Software Reliability, pp. 217-235, Elsevier, September 1990.

[Sha88]
Richard J. Shavelson
Statistical Reasoning for the Behavioral Sciences
2. ed., 1988, Allyn & Bacon, Inc., ISBN 0-205-11765-1.

7. Experimenteller Entwurf (QIP 3)

7.1 Vorbedingungen

Für die Ausführung der Aktivitäten in diesem Schritt werden die GQM-Pläne (je einer pro experimentelles Ziel) sowie ergänzend die Charakterisierung des vorhandenen Wissens (aus dem 1. QIP-Schritt) benötigt und sollten deshalb vor Ausführung dieses Schrittes existieren.

7.2 Vorgehen

Die folgende Tabelle faßt die wichtigsten Aspekte dieses QIP-Schrittes zusammen:

QIP3: Entwurf des Experiments		
Variablen	Entwurf	Instrumentierung
abhängige unabhängige	Zwei-Gruppen-Entwurf Zwei-Gruppen-Entwurf mit Vortest (Fraktional-)Faktorieller Entwurf Andere Entwürfe	Meßplan

Tabelle 6: Schritt 3 des QIP: Entwurf des Experiments

Die erste Aktivität innerhalb dieses Schrittes ist die Identifikation der abhängigen und der unabhängigen Variablen[1]. Eine unabhängige Variable ist eine Variable, die vom Experimentator manipuliert oder beobachtet werden kann, um ihre Beziehung zu anderen Variablen festzustellen [Sha88]. Abhängige Variablen sind Variablen, die in Abhängigkeit der unabhängigen Variablen gemessen werden. Bei Experimenten wird erwartet, daß sich eine Änderung der unabhängigen Variablen in einer Änderung der abhängigen Variablen widerspiegelt. Die unabhängigen und abhängigen Variablen lassen sich aus den Maßen des GQM-Plans des Experiments entnehmen. Sie spezifizieren, was im Experiment kontrolliert und was gemessen werden muß.

Die nächste Aktivität besteht aus der Erstellung des Entwurfs. Durch den Entwurf wird festgelegt, wie die Beziehung zwischen unabhängigen und abhängigen Variablen überprüft werden kann. Tabelle 6 enthält mögliche Entwürfe, die in diesem Kapitel beschrieben werden.[2] Der Entwurf ist der wichtigste und auch der schwierigste Teil der Vorbereitung eines Experiments. Im Entwurf wird beschrieben, welche Teilnehmer (einzeln oder in Teams) im Experiment welche Technik, Methode oder Werkzeug, wie und zu welchem Zeitpunkt anwenden. Dabei sollten die Eigenschaften (wie Skalierbarkeit, Wiederholbarkeit, ...) und Ziele des Experiments berücksichtigt und gleichzeitig dafür gesorgt werden, daß mit den gesammelten Daten tatsächlich die Hypothese überprüft werden kann. Die Wahl des Entwurfs hat also einen wesentlichen Einfluß auf die Validität des Experiments. Mögliche Entwürfe werden weiter unten diskutiert.

1. Schon bei der Aufstellung eines GQM-Plans kann bei der Ableitung der Maße eine Unterscheidung zwischen unabhängigen und abhängigen Variablen getroffen werden. Hier soll explizit festgehalten werden, welche Variablen abhängig und welche unabhängig sind.
2. Bei den in diesem Kapitel vorgestellten Entwürfen ist die Anzahl der Gruppen jeweils größer oder gleich zwei, d. h., die Entwürfe sind für „Replicated Project“- oder „Blocked Subject-Project“-Studien geeignet (siehe Tabelle 1, Seite 6)

Die letzte Aktivität dieses Schrittes ist die Instrumentierung des Experiments. Zur Instrumentierung sollte ein Meßplan aufgestellt werden, in dem die folgenden Aspekte beschrieben sind [Hoi94]:

- die Art und Weise, wie die Daten gesammelt werden (z. B. über Interviews),
- von wem die Daten gesammelt werden,
- wann die Daten gesammelt werden.

Alle Materialen zur Datenerfassung wie z. B. Fragebögen gehören ebenfalls zum Meßplan. Normalerweise erfolgt die Erfassung der Daten über Fragebögen, die Experimentteilnehmer im Verlauf des Experiments ausfüllen [BW84][SEL94][VC93]. Eine Erfassung von Daten kann aber auch automatisch durch ein Rechnersystem erfolgen.

Experimentelle Entwürfe

In diesem Abschnitt werden mögliche Entwürfe von Experimenten vorgestellt, mit denen Hypothesen überprüft werden können. Sie können als Ausgangspunkt für den eigenen experimentellen Entwurf dienen.

Generell gibt es zwei Möglichkeiten, Experimentteilnehmer auf Gruppen aufzuteilen [Pfl95a][Pfl95b]:

1. Blocking: Beim „Blocking“ werden Personen, die eine gemeinsame Eigenschaft haben, in einer Gruppe zusammengefaßt. Dabei muß allerdings durch die Kontrolle beim Experiment garantiert werden, daß es außer den als unabhängig spezifizierten Variablen keine weiteren Unterschiede gibt, die das Ergebnis beeinflussen können. Dies ist außerhalb eines Labors sehr schwer möglich.
2. Randomization: Bei „Randomization“ werden die zu untersuchenden Objekte oder Personen zufällig auf die verschiedenen Experimentgruppen verteilt. Gibt es außer den spezifizierten unabhängigen Variablen noch Einflüsse, so wird durch die zufällige Verteilung angenommen, daß sie in den Gruppen gleich sind und sich somit aufheben.

Im folgenden werden typische Entwürfe vorgestellt. Sie stammen aus [JSK91].

Zwei-Gruppen-Entwurf (engl.: Randomized Two-Group Design)

Bei dieser Entwurfsmethode werden die Teilnehmer zufällig auf zwei oder mehrere Gruppen verteilt (siehe Abbildung 5). Die Gruppen werden gemäß der spezifizierten unabhängigen Variablen unterschiedlich behandelt und die Ergebnisse dann verglichen. Dieser Entwurf ist die Grundlage für viele Experimente und kann vielfältig modifiziert werden. Es besteht z. B. die Möglichkeit, daß noch eine Kontrollgruppe beim Experiment teilnimmt, die keine Behandlung (z. B. Schulung in einer Technik) erfährt. Durch Kontrollgruppen ergibt sich dann die Möglichkeit, Lerneffekte beurteilen zu können. Unter einem Lerneffekt versteht man, daß die Experimentteilnehmer durch das Experiment selbst (d. h. die Übung, die sie im Verlauf des Experiments erfahren haben) bessere Ergebnisse erzielen. Lerneffekte können das Ergebnis verfälschen, sofern sie nicht beachtet werden. Für diesen Entwurf muß es mindestens zwei Gruppen geben. Durch die zufällige Zuordnung von Personen zu Gruppen wird garantiert, daß beide Gruppen vor der Behandlung annähernd gleich sind. Jeder Unterschied, der sich nach der Behandlung ergibt, kann auf die

unabhängigen Variablen zurückgeführt werden. Dieser Entwurf garantiert eine hohe Gültigkeit der Daten, da Faktoren, die die Gültigkeit gefährden könnten, durch die zufällige Zuteilung aufgehoben werden. Gültigkeit bedeutet in diesem Zusammenhang, daß sich Unterschiede in den Variablen lediglich durch die Variation der unabhängigen Variablen ergeben.

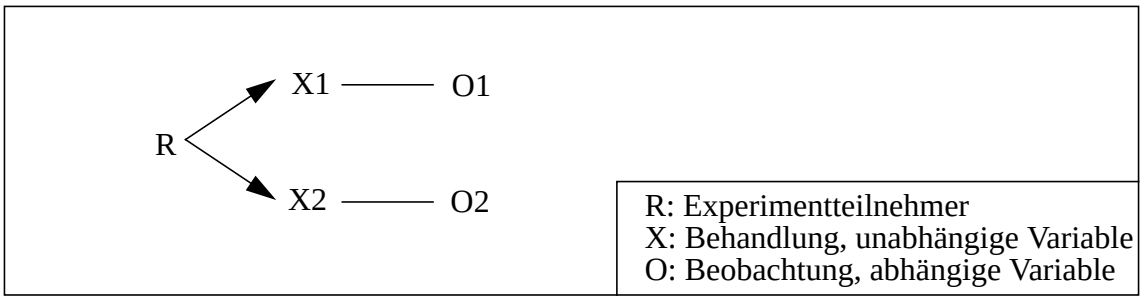

Abbildung 5: Zuordnung beim Zwei-Gruppen-Entwurf

Zwei-Gruppen-Entwurf mit Vortest (engl.: Before-After Two-Group Design)

Der Unterschied zwischen diesem Entwurf und dem Zwei-Gruppen-Entwurf liegt darin, daß die beiden Gruppen bereits vor dem eigentlichen Experiment untersucht werden. Dadurch ist eine Überprüfung möglich, ob die beiden Gruppen vor dem Experiment wirklich gleich sind, d. h., eine Kontrolle, wie zufällig die Verteilung der Eigenschaften der Teilnehmer einer Gruppe ist. Eine Variation besteht darin, vor der Aufteilung in Gruppen den Vortest durchzuführen und basierend auf den Ergebnissen, die Aufteilung in Gruppen vorzunehmen.

Desweiteren können nach der Durchführung des Experiments nicht nur die Ergebnisse der Gruppen sondern auch das einzelne Verhalten der Experimentteilnehmer vor und nach dem Experiment untersucht werden. Allerdings hat die Überprüfung auch Nachteile. Die Überprüfung vor dem Experiment kann die Ergebnisse der Experimentteilnehmer im Experiment beeinflussen, da sie für das Experiment sensibilisiert werden (Hawthorne-Effekt [Par74]). Es muß also im Einzelfall überprüft werden, ob eine Beeinflussung der Experimentteilnehmer vorliegt oder nicht.

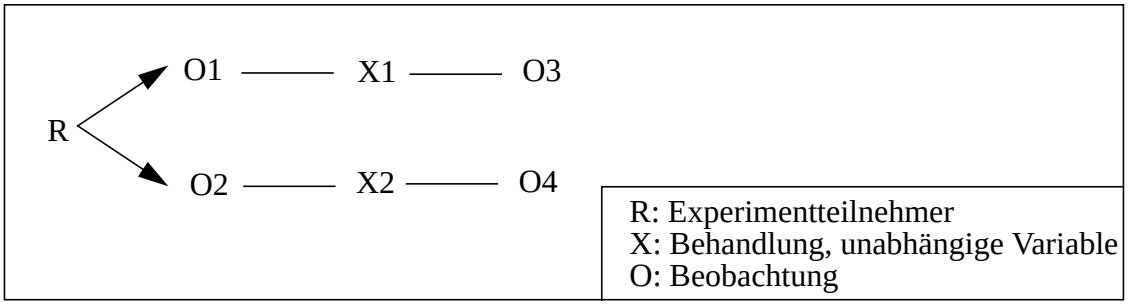

Abbildung 6: Zuordnung beim Zwei-Gruppen-Entwurf mit Vortest

Faktorieller Entwurf (engl.: Factorial Design)

Will man jetzt den Einfluß von mehreren unabhängigen Variablen messen, so werden beim faktoriellen Entwurf alle möglichen Kombinationen der unabhängigen Variablen betrachtet. Dies führt dazu, daß nicht

nur der Einfluß von den verschiedenen unabhängigen Variablen untersucht werden kann, sondern auch, wie sich die unabhängigen Variablen gegenseitig beeinflussen. Es ist dabei allerdings zu beachten, daß mit jeder neuen unabhängigen Variablen die Anzahl der Kombinationen wächst und man darauf achten muß, daß in jeder Gruppe genügend Teilnehmer sind, um eine ausreichende Menge Daten zu erhalten. Ansonsten lassen sich keine statistisch signifikanten Aussagen über Annahme oder Ablehnung einer Hypothese machen.

Wenn nicht alle Kombinationen im Entwurf berücksichtigt werden (z. B. weil einige Kombinationen nicht als sinnvoll betrachtet werden), spricht man von *fraktional-faktoriellen Entwürfen*.

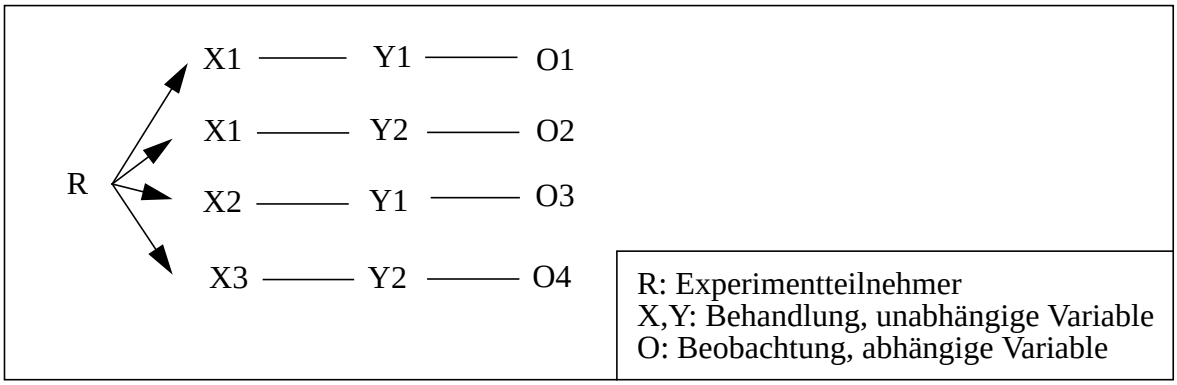

Abbildung 7: Zuordnung beim Faktor-Entwurf

Andere Entwürfe

Die bisher vorgestellten Entwürfe bilden die Grundlage für eine große Anzahl von möglichen Experimenten. Für den konkreten Fall muß überprüft werden, welcher Entwurf aufgrund der Gegebenheiten (Anzahl der Teilnehmer, Kosten, etc.) tatsächlich realisiert werden kann. Dabei kann es durchaus vorkommen, das keiner der hier vorgestellten Entwürfe geeignet ist und ein anderer Entwurf verwendet oder entwickelt werden muß.

Bei der Auswahl des Entwurfs sollte zum einen berücksichtigt werden, welche Validität die Ergebnisse haben sollen und wie die Daten ausgewertet werden können. Für eine ausführliche Beschreibung der vorgestellten Entwürfe wird auf [JSK91] verwiesen. Dort sind auch weitere Entwürfe dargestellt.

Die folgende Abbildung faßt die Aktivitäten und Produkte dieses Schrittes noch einmal zusammen:

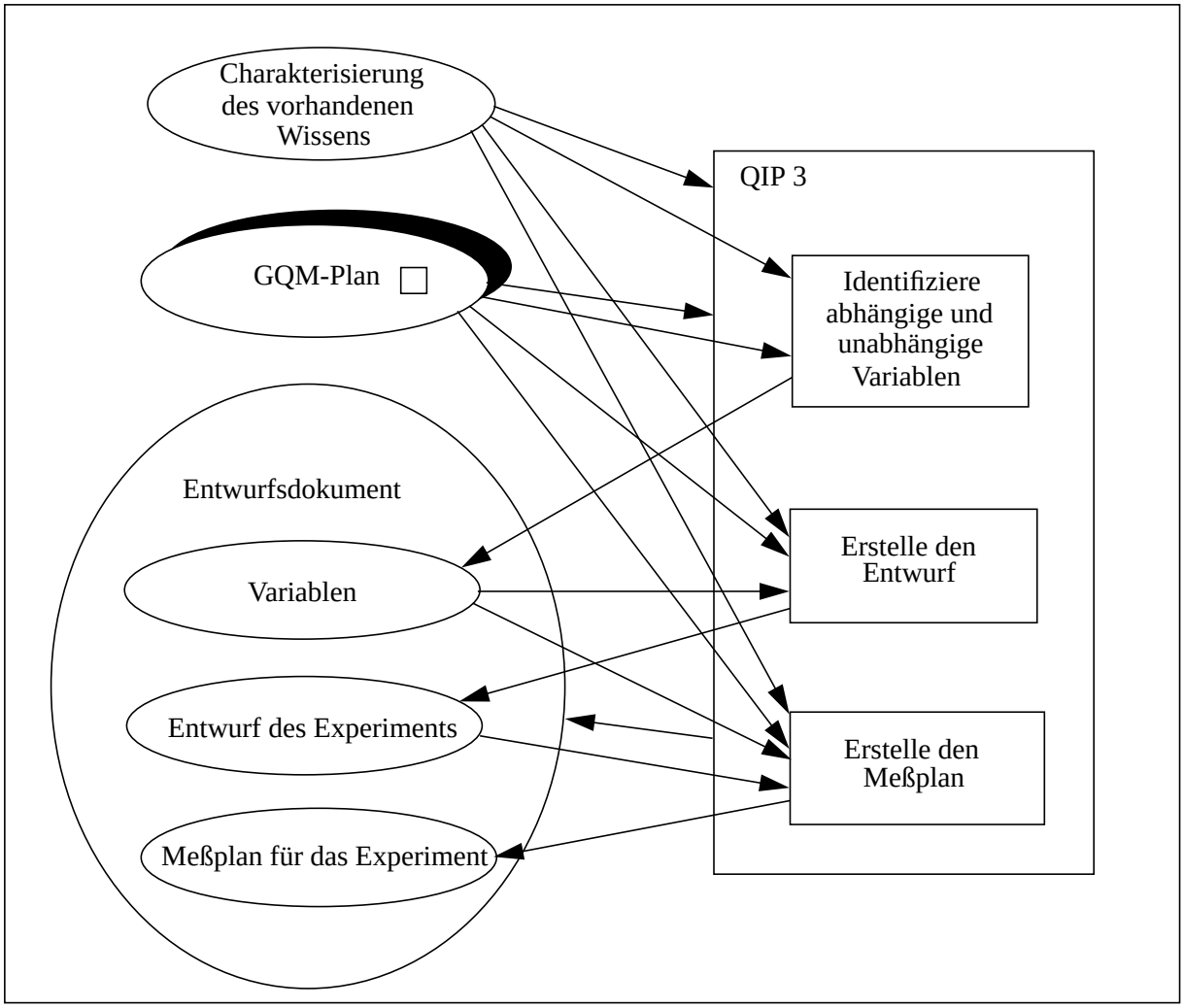

Abbildung 8: Verfeinerung des Schrittes 3 des QIP

7.3 Nachbedingungen

Nach Beendigung dieser Aktivitäten sollte ein Entwurfsdokument vorhanden sein, das eine Beschreibung der abhängigen und unabhängigen Variablen, den Entwurf des Experiments sowie eines Meßplan (incl. aller notwendigen Materialien zur Datenerfassung) umfaßt.

7.4 Beteiligte Rollen

Dieser Schritt des QIP wird vollständig durch den Experimentator durchgeführt.

7.5 Beispiel

Für ein erstes Experiment, das mit Studenten der Universität von Maryland durchgeführt wurde, ergaben sich drei *unabhängige Variablen*:

- Verwendete Testtechnik: {strukturelles Testen, funktionales Testen, Code reading},
- Software-Typ: {Textverarbeitung (P_1), abstrakte numerische Datentypen (P_2), Datenbankanwendung (P_3)},
- Kenntnisstand der Studenten: {gering, durchschnittlich, fortgeschritten}.

Die *abhängigen Variablen* waren:

- Anzahl der gefundenen Fehler,
- Prozentzahl der gefundenen Fehler,
- Zeit, die zum Finden der Fehler benötigt wurde.

Bei funktionalem und strukturellem Testen wurde noch die CPU-Zeit, die Anzahl der Durchläufe und die Anweisungsabdeckung und diverse andere abhängige Variablen untersucht.

Als *Entwurf* wurde ein fraktional-faktorieller Entwurf gewählt. Die folgende Tabelle zeigt die Zuordnung der 32 Teilnehmer zu Programm und Testtechnik. Zu beachten ist, daß jeder Teilnehmer die Testtechniken auf verschiedene Programme anwendete, um Lerneffekte zu vermeiden.

		Code reading			Funktionales Testen			Strukturelles Testen		
Teilnehmer		P_1	P_2	P_3	P_1	P_2	P_3	P_1	P_2	P_3
fortgeschritten	S_1	--	--	X	--	X	--	X	--	--
	S_2	--	X	--	X	--	--	--	--	X
	...									
	S_8	X	--	--	--	--	X	--	X	--
durchschnittlich	S_9	--	X	--	X	--	--	--	--	X
	S_{10}	--	--	X	--	X	--	X	--	--
	...									
	S_{19}	X	--	--	--	--	X	--	X	--
gering	S_{20}	--	X	--	X	--	--	--	--	X
	S_{21}	X	--	--	--	--	X	--	X	--
	...									
	S_{32}	--	--	X	--	X	--	X	--	--

Tabelle 7: Entwurf des Testexperiments

Desweiteren mußte beim Entwurf berücksichtigt werden, daß jeweils das gleiche Programm von allen Teilnehmern an einem Tag getestet wurde. Dies garantierte, daß der Informationsaustausch zwischen Teilnehmern über gefundene Fehler keinen Einfluß auf die Ergebnisse hatte, da am nächsten Tag ein Teilnehmer ein anderes Programm mit einer anderen Technik testete.

Ein Meßplan für das Experiment könnte exemplarisch folgendermaßen aussehen:

Datenerfassung: Fragebögen, in die gefundene Fehler sowie die Zeit zum Finden der Fehler eingetragen werden.

Datensammlung: Die Sammlung der Daten erfolgt durch die Experimentteilnehmer. Sie tragen die Anzahl der gefundenen Fehler sowie die dafür benötigte Zeit manuell in die Fragebögen ein. Für jede Technik wird von den Teilnehmern ein Fragebogen ausgefüllt.

Zeitpunkt: Die Daten werden im Verlauf des Experiments für jeden Teilnehmer und jede Technik gesammelt. An dieser Stelle erfolgt eine Beschreibung, welcher Experimentteilnehmer an welchem Tag welche Testtechnik anwenden soll.

7.6 Literatur

[BW84]
Victor R. Basili and D. M. Weiss
A Methodology for Collecting Valid Software Engineering Data
IEEE Transactions on Software Engineering, vol. 10, No. 5, pp. 728-738, November 1984.

[BS87]
Victor R. Basili and Richard W. Selby
Comparing the Effectiveness of Software Testing Techniques
IEEE Transactions on Software Engineering, vol. 13(12), pp. 1278-1296, December 1987.

[Hoi94]
Barbara Hoisl
A Process Model for Planning GQM-based Measurement
Technical Report STTI-94-06-E, Software-Technology-Transfer-Initiative Kaiserslautern, University of Kaiserslautern, 67653 Kaiserslautern, Germany, April 1994.

[JSK91]
Charles Judd, Eliot R. Smith, and Louise Kidder
Research Methods in Social Relations
6th ed., Harcourt Brace Jovanovich College Publishers, ISBN 0-03-031149-7.

[Par74]
H. M. Parsons
What happened at Hawthorne?
Science, vol. 183, pp. 922-932, March 1974.

[Pfl95a]
Shari L. Pfleeger
Experimental Design and Analysis in Software Engineering
Part 2: How to Set Up an Experiment
ACM SIGSOFT Software Engineering Notes, vol. 20, no. 1, January 1995.

[Pfl95b]
Shari L. Pfleeger
Experimental Design and Analysis in Software Engineering
Part 3: Types of Experimental Design
ACM SIGSOFT Software Engineering Notes, vol. 20, no. 2, March 1995.

[SEL94]
National Aeronautics and Space Administration
Software Measurement Guidebook
Tech. Rep. SEL-84-101, NASA Goddard Space Flight Center, Greenbelt MD 20771, July 1994.

[Sha88]
Richard J. Shavelson
Statistical Reasoning for the Behavioral Sciences
2. ed., 1988, Allyn & Bacon, Inc., ISBN 0-205-11765-1.

[VC93]
Jon D. Valett and Steven E. Condon
The (Mis)Use of Subjective Process Measures in Software Engineering
in Proceedings of the 18th Annual Software Engineering Workshop, pp. 161-175, NASA Goddard Space Flight Center, Greenbelt MD 20771, 1993.

8. Durchführung (QIP 4)

8.1 Vorbedingungen

Bevor das Experiment ausgeführt wird, muß der Entwurf des Experiments feststehen. Der Entwurf bestimmt letztendlich, wie die Kontrolle beim Ablauf des Experiments ausgeführt wird. Sonstige Informationen, wie, wann und von wem Daten gesammelt werden, sind dem Meßplan zu entnehmen, d. h. der Meßplan muß ebenfalls existieren. Er enthält die für die Datensammlung notwendigen Formblätter, Fragebögen, etc.

8.2 Vorgehen

Die folgende Tabelle faßt die wichtigsten Aspekte dieses QIP-Schrittes zusammen:

QIP4: Durchführung des Experiments		
Vorbereitung	Abwicklung	Messen
Pilotstudie	Experimentdurchführung	Datensammlung Validierung der Daten

Tabelle 8: Schritt 4 des QIP: Durchführung des Experiments

Wenn die Vorbedingungen für das Experiment erfüllt sind, kann eine Pilotstudie durchgeführt werden. Eine Pilotstudie kann zum Beispiel mit Mitarbeitern der eigenen Arbeitsgruppe durchgeführt werden. Die Teilnehmer einer Pilotstudie dürfen nicht die gleichen wie die Experimentteilnehmer sein. Das Ziel ist es hier nicht, Ergebnisse zu bekommen, sondern die Materialien (Fragebögen, Anleitungen oder auch Schulungen), die beim Experiment verwendet werden, auf Fehler zu untersuchen und entsprechend zu verbessern.

Das Experiment wird nach der im Entwurf spezifizierten Vorgehensweise durchgeführt. Die Daten werden entsprechend des Meßplans gesammelt. Die Meßdatenerfassung kann sowohl automatisch (durch Werkzeuge) als auch manuell (durch Formblätter, Interviews etc.) unterstützt werden. Es empfiehlt sich dringend, die Teilnehmer vor dem Experiment über Sinn und Zweck des Experiments aufzuklären. Wenn die Datensammlung durch die Experimentteilnehmer erfolgt, müssen zumindest die beim Experiment verwendeten Formulare, Frage- und Antwortbögen erklärt werden, damit von den Teilnehmern auch diejenigen Daten gesammelt werden, die gesammelt werden sollen. Dabei muß allerdings beachtet werden, daß die Teilnehmer nicht beeinflußt werden, so daß diese sich anders verhalten als angenommen.

Es muß darauf geachtet werden, daß „Kontrolle“, wie sie im Entwurf spezifiziert worden ist, auch tatsächlich ausgeübt wird, so daß nur die unabhängigen Variablen einen Einfluß auf die Ergebnisse haben. Insbesondere ist darauf zu achten, daß alle im Entwurf gemachten Annahmen erfüllt sind.

Die Validierung der experimentellen Daten stellt sicher, daß die Datensammlung korrekt erfolgt. Ihr kommt große Bedeutung zu, da die Analyseergebnisse nur so gut wie die zugrundeliegenden Daten sein können. Die Validierung der Daten kann u. a. durch klare Datenerfassungsrichtlinien, operational definierte Datenklassifikationen, redundante Fragen (Möglichkeit von Konsistenzchecks), Plausibilitätstests und nachträgliche Interviews erreicht werden.

8.3 Nachbedingungen

Die Nachbedingung für diesen Schritt des QIP besteht im Vorhandensein der ausgefüllten und überprüften Formulare, Antwortbögen, etc., d. h. generell im Vorhandensein der im Experiment gesammelten und validierten Daten.

8.4 Beteiligte Rollen

Bei der Ausführung dieser Aktivität sind sowohl der Experimentator (zur Koordination) als auch die Experimentteilnehmer beteiligt. Sofern eine Pilotstudie durchgeführt wird, sind weitere Personen beteiligt, die allerdings keine in diesem Dokument spezifizierte Rolle ausüben.

8.5 Beispiel

Das Experiment wurde wie spezifiziert durchgeführt.

8.6 Literatur

[BS87]
Victor R. Basili and Richard W. Selby
Comparing the Effectiveness of Software Testing Techniques
IEEE Transactions on Software Engineering, vol. 13(12), pp. 1278-1296, December 1987.

[BW84]
Victor R. Basili and D. M. Weiss
A Methodology for Collecting Valid Software Engineering Data
IEEE Transactions on Software Engineering, vol. 10, no. 5, pp. 728-738, November 1984.

9. Analyse und Auswertung der Daten (QIP 5)

9.1 Vorbedingungen

Die Analyse und Auswertung setzt die Existenz der im Experiment gesammelten Daten voraus. Die Daten sind in der Regel auf den im Experiment verwendeten Formularen gesammelt worden. Auch der GQM-Plan und das Entwurfsdokument werden benötigt.

9.2 Vorgehen

Die folgende Tabelle faßt die wichtigsten Aspekte dieses QIP-Schrittes zusammen:

QIP5: Auswertung und Analyse der Daten		
Analyse	Interpretation	Extrapolation
Datenüberblick	Bestätigung/Ablehnung der Hypothese(n)	Validität
Verteilung	Zweck der Studie	
Statistische Testverfahren	Forschungsfeld	
Weitere Analyseverfahren		

Tabelle 9: Schritt 5 des QIP: Auswertung und Analyse der Daten

Die bei der Durchführung des Experiments gesammelten Daten werden in diesem Schritt ausgewertet. Für die Auswertung stehen verschiedene Analyseverfahren zur Verfügung. In diesem Bericht werden statistische Testverfahren genauer vorgestellt, da diese typischerweise angewendet werden. Dabei wird/werden die Hypothese(n) mit statistischen Tests überprüft. Bevor jedoch statistische Tests angewendet werden, sollte man sich einen Überblick über die Daten verschaffen. Dabei sollte z. B. überprüft werden, ob die Daten tatsächlich so verteilt sind, wie es angenommen worden ist. Die Verteilung der Daten determiniert die Testverfahren, die zur Verfügung stehen. Es sollte derjenige Test ausgewählt werden, der die Bedingungen des Experiments am genauesten modelliert.[1] Die Anwendung der statistischen Tests auf die Daten erlaubt die Annahme oder die Ablehnung der in Schritt 2 des QIP aufgestellten Hypothesen.

Bei der Entscheidungsfindung können folgenden Fälle auftreten:

Entscheidung	Nullhypothese ist wahr	Nullhypothese ist falsch
Keine Ablehnung der Nullhypothese	richtige Entscheidung	Typ-II-Fehler
Ablehnung der Nullhypothese	Typ-I-Fehler	richtige Entscheidung

Tabelle 10: Zusammenhang zwischen Richtigkeit der Nullhypothese und Entscheidung

1. Die Auswahl eines Tests kann teilweise auch schon in der Entwurfsphase erfolgen: Wenn ein Entwurf gewählt wurde, impliziert dies bereits gewisse Testverfahren. Will man ein bestimmtes Testverfahren anwenden, so muß schon beim Entwurf darauf geachtet werden, daß der Entwurf die Bedingungen des Tests garantiert.

Anhand der Fehlerklassifikation kann das Signifikanzniveau auch als Wahrscheinlichkeit interpretiert werden, keinen Typ-I-Fehler zu begehen. Die Überprüfung, ob ein Typ-I-Fehler gemacht wurde, ist schwieriger als bei einem Typ-II-Fehler. Für die genauere Betrachtung von Typ-I-Fehlern wird auf die Literatur verwiesen (z. B. [JSK91] [Sha88]).

Der nächste Aspekt betrifft die Frage, wie sich die Ergebnisse des Experiments in die Studie über das Objekt einerseits und in das gesamte Forschungsfeld andererseits integrieren lassen.

Als letztes sollten Aussagen über die Validität des Experiments gemacht werden, wobei insbesondere Faktoren aufzuzählen sind, die möglicherweise die Validität beeinträchtigt haben.

Statistische Testverfahren

Mit Hilfe von statistischen Testverfahren wird versucht, aus den Daten, die bei einem Experiment gesammelt worden sind, gültige Schlüsse abzuleiten. Da die im Experiment gesammelten Daten lediglich eine Stichprobe darstellen, sollen die Schlußfolgerungen auch für die Gesamtheit gelten, aus der die Stichprobe gezogen wurde. Dabei sind die folgenden beiden Aspekte wichtig:

1. Welche statistische Analysetechnik wird angewendet?
2. Welchen Grad an Validität wird durch die Analysetechnik garantiert?

Generell stellt die Statistik mehrere mögliche Analysetechniken bereit, die je nach Zielrichtung verwendet werden können. Wichtige Analysetechniken sind Hypothesentests, Korrelationsanalysen und Regressionsanalysen:

1. Hypothesentests
   Bei Hypothesentests erfolgt eine formale Überprüfung, ob eine Menge von Daten die Annahme oder Ablehnung einer Hypothese rechtfertigt. Die Überprüfung der Hypothese erfolgt dabei auf der Basis von Wahrscheinlichkeiten.
2. Korrelationsanalysen
   Bei Korrelationsanalysen wird die Stärke des statistischen Zusammenhanges zwischen den zu untersuchenden Variablen untersucht.
3. Regressionsanalysen
   Regressionsanalysen versuchen ein Modell zu ermitteln, das den Zusammenhang zwischen den zu untersuchenden Variablen darstellt. Das Modell wird durch einen funktionalen Zusammenhang zwischen den betrachteten Variablen beschrieben. Sowohl bei der Korrelations- als auch bei der Regressionsanalyse werden in den meisten Fällen jeweils nur die Zusammenhänge zwischen zwei Variablen untersucht.

In diesem Bericht wird die Vorgehensweise des Experimentierens mit Hypothesentests beschrieben. Die Beschränkung auf Hypothesentests in diesem Bericht schließt allerdings nicht aus, daß die Stärke eines Zusammenhangs mit Hilfe einer Regressions- und Korrelationsanalyse untersucht wird.

Hypothesentests

Bei den Hypothesentests wird zwischen parametrischen und nichtparametrischen Tests unterschieden. Parametrische Tests erlauben gültigere Aussagen über die Hypothesen, verlangen aber auch die Erfüllung von stärkeren Vorbedingungen (u. a. die Normalverteilung der Werte, was insbesondere bei einer geringen Anzahl von Datenpunkten nicht gewährleistet werden kann). Deshalb werden im folgenden die Vor- und Nachteile von nichtparametrischen Tests kurz diskutiert. Nichtparametrische Tests haben folgende Vorteile:

1. Nichtparametrische Testverfahren eignen sich besonders, wenn die Stichprobe klein ist. Dies ist typischerweise im Software Engineering der Fall.
2. Die Annahmen, die z. B. über die Verteilung der Daten gemacht werden müssen, sind schwächer als bei parametrischen Tests.
3. Die Daten müssen nicht intervall- oder rationalskaliert sein. In manchen Fällen reichen nominalskalierte Werte aus.
4. Nichtparametrische Tests sind einfacher anzuwenden als parametrische Tests.

Diesen Vorteilen steht der Nachteil gegenüber, daß die Validität der Ergebnisse bei Anwendung von nichtparametrischen Tests geringer ist als bei parametrischen. Der Grad an Validität ist u. a. von der verwendeten statistischen Testtechnik abhängig. Validität bedeutet in diesem Zusammenhang, daß die durch die Technik getroffenen Schlußfolgerungen auf der Basis der im Experiment gesammelten Daten auch tatsächlich korrekt sind (interne Validität). Deshalb muß bei Aussagen über die Validität stets darauf geachtet werden, welche Testtechnik verwendet wurde, um Hypothesen zu überprüfen.

Beispiele für nichtparametrische Tests sind der Mann-Whitney-U-Test oder der Vorzeichentest von Wilcoxon. Da diese in den meisten Statistikbüchern (z. B. [BHH78]) erklärt sind, soll hier nicht weiter darauf eingegangen werden. Es sollte beachtet werden, daß immer das Testverfahren ausgewählt wird, das die im Entwurf des Experiments gemachten Bedingungen am genauesten modelliert.

Wurde ein entsprechender statistischer Test ausgewählt, so wird wie folgt vorgegangen:

1. Definition des Signifikanzniveaus,
2. Übernehmen der Null- und Alternativhypothese aus QIP-Schritt 2,
3. Berechnung des Testwerts (Signifikanz) aus den gesammelten Daten des QIP-Schritts 4,
4. Annahme bzw. Ablehnung der Nullhypothese in Abhängigkeit des verwendeten Testverfahrens, dem Signifikanzniveau und dem berechneten Testwert.

9.3 Nachbedingungen

Nach der Beendigung dieses Schrittes liegen die Ergebnisse des Experimentes vor. Diese bestehen insbesondere in der Aussage, ob die Hypothesen angenommen oder abgelehnt worden sind.

9.4 Beteiligte Rollen

Die Analyse der Daten erfolgt vollständig durch den Experimentator. Dieser kann die Hilfe eines Statistikers in Anspruch nehmen, sofern einer verfügbar ist.

9.5 Beispiel

Zur Überprüfung der aufgestellten Hypothesen wurde eine Varianzanalyse (ANOVA) [BHH78] durchgeführt. Als statistischer Test wurde ein F-Test verwendet. Mit Hilfe der Varianzanalyse kann man sowohl den Einfluß der unabhängigen Variablen auf die Ergebnisse einzeln als auch in der Kombination überprüfen. So kann z. B. überprüft werden, welchen Einfluß Programm 1 zusammen mit funktionalem Testen auf die Ergebnisse hat. ANOVA konnte deshalb verwendet werden, da es sich um ein „Randomized Experiment“ gehandelt hat und die Anzahl der gewonnenen Datenpunkte ausreichte, um auf Normalverteilung zu schließen. Als Signifikanzniveau wurde in der Regel 0.05 verwendet.

Für die im Beispiel von Kapitel 6 (QIP 2) aufgeführten Hypothesen ergaben sich bei der dritten Durchführung des Experiments die folgenden Ergebnisse:

Nullhypothese	Ergebnis	Signifikanzniveau
Mit strukturellem Testen werden mehr Fehler gefunden als mit funktionalem Testen	abgelehnt	0.0007
Die Fehlererkennungsrate ist bei strukturellem Testen höher als bei funktionalem Testen	abgelehnt	0.0001

Tabelle 11: Bestätigung/Ablehnung der Hypothesen aus Kapitel 6

Das Experiment umfaßte eine Reihe weiterer Hypothesen und Fragestellungen. Bei der Auswertung ergaben sich die folgenden wesentlichen Ergebnisse:

1. Bei der ersten Durchführung des Experiments mit Teilnehmern der NASA wurde mit Hilfe von Code reading mehr Fehler gefunden (absolut und relativ bezogen auf die Zeit) als mit den anderen Testtechniken.
2. Bei der ersten Studie mit Studenten der Universität von Maryland gab es keine Unterschiede zwischen Code reading und funktionalem Testen. Beide waren allerdings besser als strukturelles Testen. Im zweiten Experiment mit Studenten wurden keine Unterschiede festgestellt.
3. Bei den Studenten ergaben sich keine Unterschiede in der relativen Fehleranzahl.
4. Die Anzahl der gefundenen Fehler und die zum Finden der Fehler benötigte Zeit waren vom Programm abhängig.
5. Mit Hilfe von Code reading wurden mehr Schnittstellenfehler gefunden als mit den anderen beiden Techniken.
6. Mit Hilfe von funktionalem Testen wurden mehr Kontrollflußfehler gefunden als mit den anderen beiden Techniken.
7. In der Schätzung der Prozentzahl der gefundenen Fehler machten die Teilnehmer, die Code reading verwendet haben, die genaueste Schätzung.

Aus den Ergebnissen des Experiments läßt sich u. a. die folgende wesentliche Aussagen ableiten:

- Code reading ist mindestens so effektiv wie funktionales und strukturelles Testen, das am Computer ausgeführt werden muß.

9.6 Literatur

[BHH78]
G. E. P. Box, W. G. Hunter, and J. S. Hunter
Statistics for Experimenters
John Wiley & Sons, New York, 1978.

[BS87]
Victor R. Basili and Richard W. Selby
Comparing the Effectiveness of Software Testing Techniques
IEEE Transactions on Software Engineering, vol. 13(12), pp. 1278-1296, December 1987.

[JSK91]
Charles Judd, Eliot R. Smith, and Louise Kidder
Research Methods in Social Relations
6th ed., Harcourt Brace Jovanovich College Publishers, ISBN 0-03-031149-7.

[Pfl95c]
Shari L. Pfleeger
Experimental Design and Analysis in Software Engineering
Part 4: Choosing an Experimental Design
ACM SIGSOFT Software Engineering Notes, vol. 20, no. 3, July 1995.

[Sha88]
Richard J. Shavelson
Statistical Reasoning for the Behavioral Sciences
2. ed., 1988, Allyn & Bacon, Inc., ISBN 0-205-11765-1.

10. Know-How-Gewinn (QIP 6)

10.1 Vorbedingungen

Die Vorbedingung des letzten Schrittes besteht in der Existenz der Ergebnisse des Experiments sowie der Erfahrungen, die im Verlauf des Experiments über die ausgeführten Aktivitäten gesammelt worden sind.

10.2 Vorgehen

Die folgende Tabelle faßt die wichtigsten Aspekte dieses QIP-Schrittes zusammen:

QIP6: Sicherung der Ergebnisse und Erfahrungen		
Ergebnisse des Experiments	Erfahrungen über das Experiment	Auswirkungen
Dokumentation	Prozeßmodell für die experimentelle Vorgehensweise	Wiederholung
Hinzufügen/Erweitern/Modifizieren von:	Teilschritte	Anwendung
- Prozeßmodellen		Skalierung
- Produktmodellen		
- Qualitätsmodellen		
- GQM-Plänen		

Tabelle 12: Schritt 6 des QIP: Sicherung der Ergebnisse und Erfahrungen

Der letzte Schritt des QIP besteht in der Sicherung der Ergebnisse und Erfahrungen des Experiments. Dabei sind zwei Aspekte zu beachten:

1. Es sind die Ergebnisse zu sichern, welche die eigentliche Forschungsarbeit betreffen (insbesondere, ob die Hypothese(n) mit den im Experiment gesammelten Daten angenommen oder abgelehnt werden konnten).
2. Es sind die Ergebnisse und Erfahrungen zu sichern, die in jedem Schritt des Experiments gemacht und dokumentiert worden sind.

Die Ergebnisse des Experiments sollten dazu verwendet werden, vorhandenes Wissen über das Objekt (in Form von Prozeßmodellen, Produktmodellen, Qualitätsmodellen oder GQM-Plänen) zu erweitern bzw. zu modifizieren, um so iterativ mehr über das Objekt zu lernen. Die Dokumentation des experimentell gewonnenen Wissens über das Objekt kann dabei in einer Erfahrungsdatenbank erfolgen (sofern vorhanden). Zusätzlich können auch Folgefragen dokumentiert werden, die während des Experiments unbeantwortet blieben oder aus dem Experiment hervorgingen. Die Sicherung des Wissens schließt eine Dokumentation des Experiments, eventuell als Publikation, mit ein, wobei zu beachten ist, daß bestimmte Informationen gewissen Kriterien der Vertraulichkeit unterliegen können.

Die Sicherung der Erfahrungen über das Experiment dient in erster Linie dazu, das Prozeßmodell für die Durchführung von Experimenten bzw. der einzelnen Teilschritte zu verbessern. Auch dieses Wissen kann in der Erfahrungsdatenbank verwaltet werden.

Generell sollten auf jeden Fall die Experimententeilnehmer über die Ergebnisse informiert werden.

Die letzte Aktivität dieses Schrittes besteht darin, alle im Experiment verwendeten Dokumente, Formulare, Programme, etc. so aufzuarbeiten und zur Verfügung zu stellen, daß andere diese bei einer Wiederholung des Experiments benutzen können. Die notwendigen Informationen können dabei auch elektronisch bereitgestellt werden. Die Frage, ob die im Experiment gesammelten Daten ebenfalls verfügbar gemacht werden, hängt von den Geheimhaltungskriterien ab.

Die Auswirkungen eines Experiments können einerseits darin bestehen, daß das Experiment wiederholt und eventuell skaliert wird, um somit die Ergebnisse des ursprünglichen Experiments zu validieren und einen weiteren Beitrag zum iterativen Lernen über das Objekt zu leisten. Andererseits können die Ergebnisse eines Experiments auch Auswirkungen auf die praktische Anwendung des Objekts haben. So kann z. B. eine Testtechnik aufgrund eines experimentellen Resultats einer anderen vorgezogen werden. Dabei ist zu beachten, daß die Umsetzung der experimentellen Ergebnisse in die Praxis mit Vorsicht erfolgen muß. So können in der Praxis zusätzliche Einflußfaktoren vorhanden sein, die in einem Experiment nicht beachtet worden sind.

10.3 Nachbedingungen

Die Nachbedingung besteht aus der Existenz der gesicherten Ergebnisse und Erfahrungen aus dem Experiment, den Erfahrungen über das Experiment an sich und einer Dokumentation des Experiments.

10.4 Beteiligte Rollen

An der Ausführung des letzten Schrittes ist sowohl der Experimentator als auch der Qualitätsmanager beteiligt. Der Experimentator liefert dabei gewonnenes Wissen über das Objekt und über das Experiment an den Qualitätsmanager, der es entsprechend aufarbeitet und in die Erfahrungsdatenbank integriert.

10.5 Beispiel

Die Ergebnisse und Erfahrungen, die bei der Durchführung des Testexperiments gesammelt wurden, sind in mehreren Artikeln beschrieben bzw. referenziert worden [Sel85] [Sel93] [BS87][BSH86].

Da zu dieser Zeit noch keine Erfahrungsdatenbank vorhanden war, wurden die Ergebnisse und Erfahrungen über entsprechende Literatur gesichert. Eine Publikation der Beschreibung des Experiments erfolgte 1987 durch einen Artikel in Transactions on Software Engineering [BS87].

Die Ergebnisse des Experiments wurden zur Verbesserung des Testprozesses bei der NASA verwendet [SEL92]. Das Experiment diente also auch dazu, herauszufinden, welche der Testtechniken in der NASA-Umgebung zu bevorzugen ist. Dies bedeutet allerdings nicht, daß die Ergebnisse direkt auf andere Umgebungen übertragbar sind.

Weitere Forschungsarbeiten, die mit dem Experiment im Zusammenhang stehen, sind Studien über Fehlerklassifikation und die Verwendung insbesondere des Code readings im Cleanroom-Entwicklungsprozeß.

Das Experiment wurde nicht in einer Art und Weise dokumentiert, daß eine einfache Wiederholung möglich gewesen wäre. Dennoch wurde es 1994 wiederholt [KL95]. Dies erforderte allerdings einen direkten Erfahrungsaustausch zwischen den Experimentatoren des ersten und des zweiten Experiments.

10.6 Literatur

[BS87]
Victor R. Basili and Richard W. Selby
Comparing the Effectiveness of Software Testing Techniques
IEEE Transactions on Software Engineering, vol. 13(12), pp. 1278-1296, December 1987.

[BSH86]
Victor R. Basili, Richard W. Selby, and David H. Hutchens
Experimentation in Software Engineering
IEEE Transactions on Software Engineering, vol. SE-12, pp. 733-743, July 1986.

[KL95]
Erik Kamsties and Christopher M. Lott
An Empirical Evaluation of Three Defect-Detection Techniques
Proceedings of the 5th European Software Engineering Conference, (W. Schäfer and P. Botella, eds.), Lecture Notes in Computer Science, Sept. 1995.

[Sel85]
Richard W. Selby
Evaluations of Software Technologies: Testing, Cleanroom, and Metrics
Technical Report CS-TR-1500, Departement of Computer Science, University of Maryland, College Park, MD, 20742, May 1985.

[Sel93]
Richard W. Selby
Software Measurement and Experimentation Frameworks, Mechanisms, and Infrastructure
in Experimental Software Engineering Issues: A critical assessment and future directions, H. D. Rombach, V. R. Basili, and R. W. Selby, eds., pp. 89-106, Lecture Notes in Computer Science Nr. 706, Springer-Verlag, September 1993.

[SEL92]
National Aeronautics and Space Administration
Recommended Approach to Software Development
Revision 3, Tech. Rep. SEL-81-305, NASA Goddard Space Flight Center, Greenbelt MD 20771, June 1992.

11. Zusammenfassung und Ausblick

Dieser Bericht beschreibt ein Prozeßmodell für die Durchführung von Experimenten im Bereich Software Engineering. Das Prozeßmodell orientiert sich an den Schritten des Quality Improvement Paradigmas (QIP). Die Anwendung des QIP auf die Durchführung von Experimenten ermöglicht einen iterativen Lernprozeß. Der Lernprozeß beinhaltet, wie ein Experiment vorzubereiten, zu planen, durchzuführen, zu analysieren und zu dokumentieren ist, damit die Schlußfolgerungen, die aus dem Experiment gezogen werden, korrekt und für Dritte nachvollziehbar sind. Deshalb ist die Verwendung des Prozeßmodells insbesondere für diejenigen Personen eine Hilfe, die sich bisher nicht oder nur wenig mit der Durchführung von Experimenten befaßt haben. Der durch das Prozeßmodell beschriebene Rahmen kann dabei sowohl allgemein für die Durchführung von Experimenten innerhalb des Software Engineerings als auch speziell für die Durchführung von Experimenten innerhalb des SFB 501 angewendet werden.
Anhand der einzelnen Schritte des QIP werden die für die Durchführung von Experimenten notwendigen Aktivitäten und Ergebnisse beschrieben. In einem konkreten Experiment können die einzelnen Schritte dann instantiiert und konkretisiert werden.

Die weitere Arbeit an diesem Vorgehensmodell beinhaltet

1. die Verifikation und Validation der Vorgehensweise bei der Durchführung von Experimenten und die darauf basierende Verbesserungen des Vorgehensmodells;
2. die weitere Verfeinerung der einzelnen Schritte des QIP. Dabei soll insbesondere auf den Aspekt eingegangen werden, wie ein Schritt konkret auszuführen ist.

Das Ziel ist es, nicht nur mehr über Software-Entwicklungsprozesse und -produkte zu erfahren, sondern auch über die Art und Weise, wie Experimente im Software Engineering durchgeführt werden können.

Danksagung

Die Autoren bedanken sich bei Dipl.-Inform. Christiane Differding, Dipl.-Inform. Frank Kollnischko, Dipl.-Inform. Martin Verlage und Dipl.-Inform. Stefan Vorwieger für ihre Hinweise bei der Durchsicht des Manuskripts.

Anhang A: Definitionen

Die folgenden Definitionen dienen der kurzen Erläuterung einiger im Dokument verwendeten Begriffe. Dabei werden folgende Aspekte berücksichtigt:

1. Es werden die wesentlichen Begriffe, die zur Beschreibung des Vorgehensmodells dienen, beschrieben.
2. Es werden keine Begriffe erklärt, die konkrete Techniken (z. B.: Mann-Whitney-U-Test) beschreiben.
3. Es werden Begriffe, hinter denen sich umfassende Konzepte verbergen, kurz erläutert und die entsprechende Literatur referenziert.

Abhängige Variable (engl.: Dependent variable)

Charakteristik, Attribut oder Eigenschaft einer Person oder Objekts, die beim Experiment gemessen werden kann.

Alternativhypothese (engl.: Alternative hypothesis)

Die Alternativhypothese bildet den Gegensatz zur Nullhypothese. Die Alternativhypothese stimmt mit der experimentellen Hypothese überein.

Erfahrungsdatenbank (engl.: Experience base)

Datenbank innerhalb der Experience Factory, in der bisher gemachte Erfahrungen und bereits vorhandenes Wissen abgelegt ist.

(siehe [SEL92])

Experience Factory

Organisationsstruktur zur Unterstützung von kontinuierlicher Verbesserung (QIP) von Softwareentwicklungsprozessen.

(siehe [BCR94a], [BM95])

Experiment (engl.: Experiment)

Unter einem Experiment versteht man eine Studie zur Überprüfung von Hypothesen. Dabei können die wichtigsten Faktoren, die einen Einfluß auf die Ergebnisse haben, kontrolliert und gemessen werden.

Experimentelle Hypothese (engl.: Experimental hypothesis)

Eine experimentelle Hypothese oder kurz Hypothese ist eine vorläufige Behauptung über eine Beziehung zwischen zwei oder mehreren Faktoren. Die Behauptung ist vorläufig, da sie noch experimentell überprüft werden muß. Dazu werden aus der experimentellen Hypothese die Nullhypothese und die Alternativhypothese abgeleitet, die anhand der im Experiment gesammelten Daten statistisch überprüft werden können.

Goal/Question/Metric-Ansatz (GQM)

Paradigma zur Unterstützung von zielorientiertem Messen und Bewerten. Ausgehend von Zielen werden Fragen gestellt, die das Ziel charakterisieren. Die Beantwortung der Fragen erfolgt mit Maßen (siehe [Bas92], [BCR94b], [Hoi94], [Rom90]).

Korrelationsanalyse (engl.: Correlation analysis)

Bei Korrelationsanalysen wird die Stärke des statistischen Zusammenhanges zwischen den zu untersuchenden Variablen untersucht.

Maß (engl.: Measure)

Ein Maß ist eine Abbildung zwischen Objekten des Software Engineerings und Objekten aus dem Bereich der Mathematik [BCR94b]. Objekte des Software Engineerings können Produkte, Prozesse oder auch Projekte sein. Objekte aus dem Bereich der Mathematik können Zahlen oder Vektoren sein. Die Abbildung kann auf verschiedenen Skalen wie z. B. einer Nominalskala definiert sein.

Methode (engl.: Method)

Unter einer Methode wird eine planmäßig anwendbare und begründete Technik verstanden, mit der vorgegebene Ziele erreicht werden können. Eine Methode beschreibt sowohl, was und wie etwas gemacht wird, als auch, unter welchen Umständen eine Methode wie gut ist.

Meßplan (engl.: Measurement plan)

Unter einem Meßplan versteht man eine genaue Beschreibung der Art und Weise, wie Daten gesammelt werden [Hoi94]. Im Meßplan sollte festgelegt sein,

- wie Daten gesammelt werden,
- von wem Daten gesammelt werden,
- wann welche Daten gesammelt werden.

Desweiteren gehören alle Materialien zum Meßplan, mit denen die Daten gesammelt werden sollen.

Nullhypothese (engl.: Null hypothesis)

Die Nullhypothese ist eine ganz bestimmte Annahme über die Beziehung zwischen zwei oder mehreren Variablen oder der Verteilung der Daten einer Variablen. Aufgrund der tatsächlich beobachteten Daten beim Experiment soll entschieden werden, ob die Nullhypothese abzulehnen ist oder nicht. Die Nichtablehnung einer Nullhypothese bedeutet nicht, daß diese richtig ist!

Produkt/ Produktmodell (engl.: Product/ Product model)

Als Produkt wird jede Informationseinheit im Rahmen des Software Engineerings bezeichnet. Ein Produktmodell beschreibt Charakteristika einer Klasse von Produkten.

Prozeß/ Prozeßmodell (engl.: Process/ Process model)

Als Prozeß wird jede Aktivität im Rahmen des Software Engineerings bezeichnet. Ein Prozeßmodell beschreibt Charakteristika einer Klasse von Prozessen.

Quality Improvement Paradigm (QIP)

Paradigma zur kontinuierlichen Verbesserung innerhalb der Software-Entwicklung. Das QIP besteht aus einem Prozeß, der sowohl Schritte für die Planung, Ausführung und Beurteilung von Softwareentwicklungsprojekten als auch für die Verwendung von bereits gemachten Erfahrungen und die Sicherung von neuen Erfahrungen enthält.

(siehe [BC93])

Qualität/ Qualitätsmodell (engl.: Quality/ Quality model)

Als Qualität wird jede Eigenschaft eines Produktes oder Prozesses bezeichnet. Ein Qualitätsmodell beschreibt eine Qualitätseigenschaft für eine Klasse von Kontextcharakteristika.

Regressionsanalyse (engl.: Regression analysis)

Regressionsanalysen versuchen ein Modell zu ermitteln, das den Zusammenhang zwischen den zu unter suchenden Variablen darstellt. Das Modell wird durch einen funktionalen Zusammenhang zwischen den betrachteten Variablen beschrieben.

Replikation (engl.: Replication)

Unter Replikation wird die Wiederholung eines Experiments unter den gleichen Bedingungen verstanden. Insbesondere ist bei einer Replikation eines Experiments darauf zu achten, daß Kontrolle in der selben Art und Weise ausgeführt wird, wie beim ursprünglichen Experiment.

Signifikanzniveau (engl.: Level of significance)

Die Ablehnung einer richtigen Nullhypothese heißt Fehler 1. Art. Die Wahrscheinlichkeit für die Ablehnung einer richtigen Nullhypothese heißt Signifikanzniveau.

Statistischer Test (engl.: Statistical test)

Ein statistischer Test ist ein Verfahren zur Überprüfung von Hypothesen, d. h. von Annahmen über die Beziehungen zwischen abhängigen und unabhängigen Variablen oder Verteilungen, aufgrund der gesammelten Daten. Die Anwendung eines statistischen Tests führt zur Annahme oder Ablehnung der Nullhypothese.

Technik (engl.: Technique)

Unter einer Technik wird eine Vorschrift zur Durchführung einer bestimmten Aktivität verstanden. Eine Technik beschreibt, was und wie etwas gemacht wird.

Unabhängige Variable (engl.: Independent variable / State variable)

Charakteristik, Attribut oder Eigenschaft einer Person oder Objekts, die beim Experiment beobachtet oder manipuliert wird, um ihren Einfluß auf die abhängige(n) Variable(n) zu überprüfen.

Validität (engl.: Validity)

Unter Validität versteht man zum einen die Tatsache, daß die Ergebnisse des Experiments tatsächlich korrekt und nicht durch Fehler entstanden sind und zum anderen, inwieweit die Ergebnisse des Experiments in einem anderen Kontext verwendet werden können.

Werkzeug (engl.: Tool)

Unter einem Werkzeug wird eine computergestützte Technik oder Methode verstanden.

Literaturverzeichnis

[Bas92]
Victor R. Basili
Software Modeling and Measurement: The Goal/Question/Metric Paradigm
Technical Report CS-TR-2956, Department of Computer Science, University of Maryland, College Park, MD, 20742, September 1992.

[Bas93]
Victor R. Basili
The Experimental Paradigm in Software Engineering
in H. D. Rombach, V. R. Basili, and R. W. Selby, editors, Experimental Software Engineering Issues: A critical assessment and future directions, page 3-12, Lecture Notes in Computer Science Nr. 706, Springer-Verlag, September 1993.

[Bas95]
Victor R. Basili
The Experience Factory and its relationship to other quality approaches
In Marvin Zelkowitz, editor, Advances in Computers, vol. 41, Seiten 65-82, Academic Press, 1995.

[BCR94a]
Victor R. Basili, Gianluigi Caldiera, and H. Dieter Rombach
The Experience Factory
Encyclopedia of Software Engineering (John J. Marciniak, ed.), vol. 1, pp. 469-476,
John Wiley & Sons, 1994.

[BCR94b]
Victor R. Basili, Gianluigi Caldiera, and H. Dieter Rombach
The Goal Question Metric Approach
Encyclopedia of Software Engineering (John J. Marciniak, ed.), vol. 1, pp. 528-532,
John Wiley & Sons, 1994.

[BD+95]
Alfred Bröckers, Christiane Differding, Barbara Hoisl, Frank Kollnischko, Christopher M. Lott, Jürgen Münch, Martin Verlage, and Stefan Vorwieger
A graphical representation schema for the software process modeling language MVP-L
Interner Bericht Nr. 270/95, Fachbereich Informatik, Universität Kaiserslautern, June 1995.

[BHH78]
G. E. P. Box, W. G. Hunter, and J. S. Hunter
Statistics for Experimenters
John Wiley & Sons, New York, 1978.

[BM95]
Victor R. Basili and Frank McGarry
The Experience Factory: How to Built and Run One
Tutorial M1, 17th International Conference on Software Engineering, Seattle, Washington, USA, April 1995.

[BS87]
Victor R. Basili and Richard W. Selby
Comparing the Effectiveness of Software Testing Techniques
IEEE Transactions on Software Engineering, vol. 13(12), pp. 1278-1296, December 1987.

[BSH86]
Victor R. Basili, Richard W. Selby, and David H. Hutchens
Experimentation in Software Engineering
IEEE Transactions on Software Engineering, vol. SE-12, pp. 733-743, July 1986.

[BW84]
Victor R. Basili and D. M. Weiss
A Methodology for Collecting Valid Software Engineering Data
IEEE Transactions on Software Engineering, vol. 10, No. 5, pp. 728-738, November 1984.

[Cur80]
Bill Curtis
Measurement and Experimentation in Software Engineering
Proceedings of the IEEE, vol. 68, pp. 1144-1157, September 1980.

[Fen91]
Norman E. Fenton
Software Metrics: A Rigorous Approach
Chapman & Hall, London, 1991.

[Fen94]
Norman E. Fenton
Software Measurement: A Necessary Scientific Basis
IEEE Transactions on Software Engineering, vol. 20, nr. 3, pp. 199-206,
March 1994.

[Gla94]
Robert L. Glass
The Software-Research Crisis
IEEE Software, pp. 42-47, November 1994.

[Hoi94]
Barbara Hoisl
A Process Model for Planning GQM-based Measurement
Technical Report STTI-94-06-E, Software-Technology-Transfer-Initiative Kaiserslautern,
University of Kaiserslautern, 67653 Kaiserslautern, Germany, April 1994.

[JSK91]
Charles Judd, Eliot R. Smith, and Louise Kidder
Research Methods in Social Relations
6th ed., Harcourt Brace Jovanovich College Publishers, ISBN 0-03-031149-7.

[Kit96a]
Barbara A. Kitchenham
Evaluating Software Engineering Methods and Tools
Part 1: The Evaluation Context and Evaluation Methods
ACM SIGSOFT Software Engineering Notes, vol. 21, no. 1, January 1996.

[Kit96b]
Barbara A. Kitchenham
Evaluating Software Engineering Methods and Tools
Part 2: Selecting an Appropriate Evaluation Method - Technical Criteria
ACM SIGSOFT Software Engineering Notes, vol. 21, no. 2, March 1996.

[KL95]
Erik Kamsties and Christopher M. Lott
An Empirical Evaluation of Three Defect-Detection Techniques
Proceedings of the 5th European Software Engineering Conference, (W. Schäfer
and P. Botella, eds.), Lecture Notes in Computer Science, Sept. 1995.

[Mar95]
Marco van Maris
GQM-DIVA: Ein Werkzeug zur Definition, Interpretation und Validation von GQM-Plänen
Diplomarbeit, Universität Kaiserslautern, Fachbereich Informatik, Mai 1995.

[NAS94]
National Aeronautics and Space Administration (NASA)
Software Engineering Program: Profile of Software at the Goddard Space Flight Center
Nasa-Report NASA-RPT-002-94, Washington DC, April 1994.

[Par74]
H. M. Parsons
What happened at Hawthorne?
Science, vol. 183, pp. 922-932, March 1974.

[Pfl94]
Shari L. Pfleeger
Experimental Design and Analysis in Software Engineering
Part 1: The Language of Case Studies and Formal Experiments
ACM SIGSOFT Software Engineering Notes, vol. 19, no. 4, October 1994.

[Pfl95a]
Shari L. Pfleeger
Experimental Design and Analysis in Software Engineering
Part 2: How to Set Up an Experiment
ACM SIGSOFT Software Engineering Notes, vol. 20, no. 1, January 1995.

[Pfl95b]
Shari L. Pfleeger
Experimental Design and Analysis in Software Engineering
Part 3: Types of Experimental Design
ACM SIGSOFT Software Engineering Notes, vol. 20, no. 2, March 1995.

[Pfl95c]
Shari L. Pfleeger
Experimental Design and Analysis in Software Engineering
Part 4: Choosing an Experimental Design
ACM SIGSOFT Software Engineering Notes, vol. 20, no. 3, July 1995.

[Pfl95d]
Shari L. Pfleeger
Experimental Design and Analysis in Software Engineering
Part 5: Analyzing the Data
ACM SIGSOFT Software Engineering Notes, vol. 20, no. 5, December 1995.

[Rom90]
H. Dieter Rombach
Practical Benefits of Goal-oriented Measurement
Proceedings of the Annual Workshop of the Centre for Software Reliability, pp. 217-235, Elsevier, September 1990.

[Rom93]
H. Dieter Rombach, Victor R. Basili, and Richard W. Selby, eds.
Experimental Software Engineering Issues: A critical assessment and future directions
Lecture Notes in Computer Science Nr. 706, Springer-Verlag, September 1993.

[Sel85]
Richard W. Selby
Evaluations of Software Technologies: Testing, Cleanroom, and Metrics
Technical Report CS-TR-1500, Departement of Computer Science, University of Maryland, College Park, MD, 20742, May 1985.

[Sel93]
Richard W. Selby
Software Measurement and Experimentation Frameworks, Mechanisms, and Infrastructure
in [Rom93], pp. 89-106.

[SEL92]
National Aeronautics and Space Administration
Recommended Approach to Software Development
Revision 3, Tech. Rep. SEL-81-305, NASA Goddard Space Flight Center, Greenbelt MD 20771, June 1992.

[SEL94]
National Aeronautics and Space Administration
Software Measurement Guidebook
Tech. Rep. SEL-84-101, NASA Goddard Space Flight Center, Greenbelt MD 20771, July 1994.

[Sha88]
Richard J. Shavelson
Statistical Reasoning for the Behavioral Sciences
2. ed., 1988, Allyn & Bacon, Inc., ISBN 0-205-11765-1.

[SFB94]
Sonderforschungsbereich 1496 [neue Nr.: 501]
Entwicklung großer Systeme mit generischen Methoden
Finanzierungsantrag 1995 - 1996 - 1997, Universität Kaiserslautern, Fachbereich Informatik, Juni 1994.

[VC93]
Jon D. Valett and Steven E. Condon
The (Mis)Use of Subjective Process Measures in Software Engineering
in Proceedings of the 18th Annual Software Engineering Workshop, pp. 161-175, NASA Goddard Space Flight Center, Greenbelt MD 20771, 1993.