HUMAN-COMPUTER INTERACTION (IS252) CHAPTER SEVEN

M.M EMAN M. HASAN

FIRST SEMESTER 2019-2020

CHAPTER 7: INTRODUCING EVALUATION

7.1 INTRODUCTION

7.2 WHAT, WHY, AND WHEN TO EVALUATE

7.2.1 WHAT TO EVALUATE

7.2.2 WHY YOU NEED TO EVALUATE

7.2.3 WHEN TO EVALUATE

7.3 HUTCH WORLD CASE STUDY

7.3.1 HOW THE TEAM GOT STARTED: EARLY DESIGN IDEAS

7.3.2 HOW WAS THE TESTING DONE?

7.3.3 WAS IT TESTED AGAIN?

7.3.4 LOOKING TO THE FUTURE

7.4 EVALUATION PARADIGMS AND TECHNIQUES

7.4.1 EVALUATION PARADIGMS

7.4.2 TECHNIQUES

7.5 DECIDE: A FRAMEWORK TO GUIDE EVALUATION

7.6 DISCUSSION

7.1 INTRODUCTION

- THIS CHAPTER BEGINS BY DISCUSSING ***WHAT*** EVALUATION IS, ***WHY*** EVALUATION IS IMPORTANT, AND ***WHEN*** TO USE DIFFERENT EVALUATION TECHNIQUES AND APPROACHES. THEN A CASE STUDY IS PRESENTED ABOUT THE EVALUATION TECHNIQUES USED BY MICROSOFT RESEARCHERS AND THE FRED HUTCHINSON CANCER RESEARCH CENTER IN DEVELOPING HUTCHWORLD (CHENG ET AL., 2000), A VIRTUAL WORLD TO SUPPORT CANCER PATIENTS, THEIR FAMILIES, AND FRIENDS. THIS CASE STUDY IS CHOSEN BECAUSE IT ILLUSTRATES HOW A RANGE OF TECHNIQUES IS USED DURING THE DEVELOPMENT OF A NEW PRODUCT. IT INTRODUCES SOME OF THE PRACTICAL PROBLEMS THAT EVALUATORS ENCOUNTER AND SHOWS HOW ITERATIVE PRODUCT DEVELOPMENT IS INFORMED BY A SERIES OF EVALUATION STUDIES. THE HUTCH WORLD STUDY ALSO LAYS THE FOUNDATION FOR THE EVALUATION FRAMEWORK. THE MAIN AIMS OF THIS CHAPTER ARE TO:
    - EXPLAIN THE KEY CONCEPTS AND TERMS USED TO DISCUSS EVALUATION.
    - DISCUSS AND CRITIQUE THE HUTCHWORLD CASE STUDY.
    - EXAMINE HOW DIFFERENT TECHNIQUES ARE USED AT DIFFERENT STAGES IN THE DEVELOPMENT OF HUTCHWORLD.
- SHOW HOW DEVELOPERS COPE WITH REAL-WORLD CONSTRAINTS IN THE DEVELOPMENT OF HUTCHWORLD.

7.2 WHAT, WHY, AND WHEN TO EVALUATE

- USERS WANT SYSTEMS THAT ARE EASY TO LEARN AND TO USE AS WELL AS EFFECTIVE, EFFICIENT, SAFE, AND SATISFYING. BEING ENTERTAINING, ATTRACTIVE, AND CHALLENGING, ETC. IS ALSO ESSENTIAL FOR SOME PRODUCTS. ***SO***, KNOWING WHAT TO EVALUATE, WHY IT IS IMPORTANT, AND WHEN TO EVALUATE ARE KEY SKILLS FOR INTERACTION DESIGNERS.

7.2.1 WHAT TO EVALUATE

- THERE IS A HUGE VARIETY OF INTERACTIVE PRODUCTS WITH A VAST ARRAY OF FEATURES THAT NEED TO BE EVALUATED. SOME FEATURES, SUCH AS THE SEQUENCE OF LINKS TO BE FOLLOWED TO FIND AN ITEM ON A WEBSITE, ARE OFTEN BEST EVALUATED IN A LABORATORY, SINCE SUCH A SETTING ALLOWS THE EVALUATORS TO CONTROL WHAT THEY WANT TO INVESTIGATE. OTHER ASPECTS, SUCH AS WHETHER A COLLABORATIVE TOY IS ROBUST AND WHETHER CHILDREN ENJOY INTERACTING WITH IT, ARE BETTER EVALUATED IN NATURAL SETTINGS, SO THAT EVALUATORS CAN SEE WHAT CHILDREN DO WHEN LEFT TO THEIR OWN DEVICES.

7.2.2 WHY YOU NEED TO EVALUATE

JUST AS DESIGNERS SHOULDN'T ASSUME THAT EVERYONE IS LIKE THEM, THEY ALSO SHOULDN'T PRESUME THAT FOLLOWING DESIGN GUIDELINES GUARANTEES GOOD USABILITY, EVALUATION IS NEEDED TO CHECK THAT USERS CAN USE THE PRODUCT AND LIKE IT.

TOGNAZZI POINTS OUT THAT THERE ARE FIVE GOOD REASONS FOR INVESTING IN USER TESTING:

1. PROBLEMS ARE FIXED BEFORE THE PRODUCT IS SHIPPED, NOT AFTER.
2. THE TEAM CAN CONCENTRATE ON REAL PROBLEMS, NOT IMAGINARY ONES.
3. ENGINEERS CODE INSTEAD OF DEBATING.
4. TIME TO MARKET IS SHARPLY REDUCED.
5. FINALLY, UPON FIRST RELEASE, YOUR SALES DEPARTMENT HAS A ROCK-SOLID DESIGN IT CAN SELL WITHOUT HAVING TO PEPPER THEIR PITCHES WITH HOW IT WILL ALL ACTUALLY WORK IN RELEASE 1.1 OR 2.0.

7.2.3 WHEN TO EVALUATE

- THE PRODUCT BEING DEVELOPED MAY BE A BRAND-NEW PRODUCT OR AN UPGRADE OF AN EXISTING PRODUCT. IF THE PRODUCT IS NEW, THEN CONSIDERABLE TIME IS USUALLY INVESTED IN MARKET RESEARCH. DESIGNERS OFTEN SUPPORT THIS PROCESS BY DEVELOPING MOCKUPS OF THE POTENTIAL PRODUCT THAT ARE USED TO ELICIT REACTIONS FROM POTENTIAL USERS. AS WELL AS HELPING TO ASSESS MARKET NEED, THIS ACTIVITY CONTRIBUTES TO UNDERSTANDING USERS' NEEDS AND EARLY REQUIREMENTS UATION IS TO ASSESS HOW WELL A DESIGN FULFILLS USERS' NEEDS AND WHETHER USERS LIKE IT.
- IN THE CASE OF AN UPGRADE, THERE IS LIMITED SCOPE FOR CHANGE AND ATTENTION IS FOCUSED ON IMPROVING THE OVERALL PRODUCT. THIS TYPE OF DESIGN IS WELL SUITED TO USABILITY ENGINEERING IN WHICH EVALUATIONS COMPARE USER PERFORMANCE AND ATTITUDES WITH THOSE FOR PREVIOUS VERSIONS. SOME PRODUCTS, SUCH AS OFFICE SYSTEMS, GO THROUGH MANY VERSIONS, AND SUCCESSFUL PRODUCTS MAY REACH DOUBLE DIGIT VERSION NUMBERS. IN CONTRAST, NEW PRODUCTS DO NOT HAVE PREVIOUS VERSIONS AND THERE MAY BE NOTHING COMPARABLE ON THE MARKET, SO MORE RADICAL CHANGES ARE POSSIBLE IF EVALUATION RESULTS INDICATE A PROBLEM.
- EVALUATIONS DONE DURING DESIGN TO CHECK THAT THE PRODUCT CONTINUES TO MEET USERS' NEEDS ARE KNOW AS *FORMATIVE EVALUATIONS*. EVALUATIONS THAT ARE DONE TO ASSESS THE SUCCESS OF A FINISHED PRODUCT, SUCH AS THOSE TO SATISFY A SPONSORING AGENCY OR TO CHECK THAT A STANDARD IS BEING UPHELD, ARE KNOW AS *SUMMATIVE EVALUATION*. AGENCIES SUCH AS NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY (NIST) IN THE USA, THE INTERNATIONAL STANDARDS ORGANIZATION (ISO) AND THE BRITISH STANDARDS INSTITUTE (BSI) SET STANDARDS BY WHICH PRODUCTS PRODUCED BY OTHERS ARE EVALUATED.

7.3 HUTCHWORLD CASE STUDY

- HUTCHWORLD IS A DISTRIBUTED VIRTUAL COMMUNITY DEVELOPED THROUGH COLLABORATION BETWEEN MICROSOFT'S VIRTUAL WORLDS RESEARCH GROUP AND LIBRARIANS AND CLINICIANS AT THE FRED HUTCHINSON CANCER RESEARCH CENTER IN SEATTLE, WASHINGTON. THE SYSTEM ENABLES CANCER PATIENTS, THEIR CAREGIVERS, FAMILY, AND FRIENDS TO CHAT WITH ONE ANOTHER, TELL THEIR STORIES, DISCUSS THEIR EXPERIENCES AND COPING STRATEGIES, AND GAIN EMOTIONAL AND PRACTICAL SUPPORT FROM ONE ANOTHER (CHENG ET. AL.,2000). THE DESIGN TEAM DECIDED TO FOCUS ON THIS PARTICULAR POPULATION BECAUSE CAREGIVERS AND CANCER PATIENTS ARE SOCIALLY ISOLATED: CANCER PATIENTS MUST OFTEN AVOID PHYSICAL CONTACT WITH OTHERS BECAUSE THEIR TREATMENTS SUPPRESS THEIR IMMUNE SYSTEMS. SIMILARLY, THEIR CAREGIVERS HAVE TO BE CAREFUL NOT TO TRANSMIT INFECTIONS TO PATIENTS.

7.3.1 HOW THE DESIGN TEAM GOT STARTED: EARLY DESIGN IDEAS

- BEFORE DEVELOPING THIS PRODUCT, THE TEAM NEEDED TO LEARN ABOUT THE PATIENT EXPERIENCE AT THE FRED HUTCHINSON CENTER. FOR INSTANCE, WHAT IS THE TYPICAL TREATMENT PROCESS, WHAT RESOURCES ARE AVAILABLE TO THE PATIENT COMMUNITY, AND WHAT ARE THE

- NEEDS OF THE DIFFERENT USER GROUPS WITHIN THIS COMMUNITY? THEY HAD TO BE PARTICULARLY CAREFUL ABOUT DOING THIS BECAUSE MANY PATIENTS WERE VERY SICK. CANCER PATIENTS ALSO TYPICALLY GO THROUGH BOUTS OF LOW EMOTIONAL AND PHYSICAL ENERGY.

- CAREGIVERS ALSO MAY HAVE DIFFICULT EMOTIONAL TIMES, INCLUDING DEPRESSION, EXHAUSTION, AND STRESS. FURTHERMORE, USERS VARY ALONG OTHER DIMENSIONS, SUCH AS EDUCATION AND EXPERIENCE WITH COMPUTERS, AGE AND GENDER AND THEY COME FROM DIFFERENT CULTURAL BACKGROUNDS WITH DIFFERENT EXPECTATIONS.

- THE DEVELOPMENT TEAM DECIDED THAT HUTCHWORLD SHOULD BE AVAILABLE FOR PATIENTS ANY TIME OF DAY OR NIGHT, REGARDLESS OF THEIR GEOGRAPHICAL LOCATION. THE TEAM'S INFORMAL VISITS TO THE FRED HUTCHINSON CENTER LED TO THE DEVELOPMENT OF AN EARLY PROTOTYPE. THEY FOLLOWED A USER-CENTERED DEVELOPMENT METHODOLOGY. HAVING GOT A GOOD FEEL FOR THE USERS' NEEDS, THE TEAM BRAINSTORMED DIFFERENT IDEAS FOR AN ORGANIZING THEME TO SHAPE THE CONCEPTUAL DESIGN A CONCEPTUAL MODEL POSSIBLY BASED ON A METAPHOR. AFTER MUCH DISCUSSION, THEY DECIDED TO MAKE THE DESIGN RESEMBLE THE OUTPATIENT CLINIC LOBBY OF THE FRED HUTCHINSON CANCER RESEARCH CENTER. BY USING THIS REAL-WORLD METAPHOR, THEY HOPED THAT THE USERS WOULD EASILY INFER WHAT FUNCTIONALITY WAS AVAILABLE IN HUTCHWORLD FROM THEIR KNOWLEDGE OF THE REAL CLINIC. THE NEXT STEP WAS TO DECIDE UPON THE KIND OF COMMUNICATION ENVIRONMENT TO USE. SHOULD IT BE SYNCHRONOUS OR ASYNCHRONOUS? WHICH WOULD SUPPORT SOCIAL AND AFFECTIVE COMMUNICATIONS BEST? **A** SYNCHRONOUS CHAT ENVIRONMENT WAS SELECTED BECAUSE THE TEAM THOUGHT THAT THIS WOULD BE MORE REALISTIC AND PERSONAL THAN AN ASYNCHRONOUS ENVIRONMENT. THEY ALSO DECIDED TO INCLUDE 3D PHOTOGRAPHIC AVATARS SO THAT USERS COULD ENJOY HAVING AN IDENTIFIABLE ONLINE PRESENCE AND COULD EASILY RECOGNIZE EACH OTHER.

- THE PROTOTYPE WAS REVIEWED WITH USERS THROUGHOUT EARLY DEVELOPMENT AND WAS LATER TESTED MORE RIGOROUSLY IN THE REAL ENVIRONMENT OF THE HUTCH CENTER USING A VARIETY OF TECHNIQUES.

- A MICROSOFT PRODUCT CALLED V-CHAT WAS USED TO DEVELOP SECOND INTERACTIVE PROTOTYPE WITH THE SUBSET OF THE FEATURES IN THE PRELIMINARY DESIGN ,HOWEVER, ONLY THE LOBBY WAS FULLY DEVELOPED.

- BEFORE TESTING COULD BEGIN, THE TEAM HAD TO SOLVE SOME LOGISTICAL ISSUES. THERE WERE TWO KEY QUESTIONS. WHO WOULD PROVIDE TRAINING FOR THE TESTERS AND HELP FOR THE PATIENTS? AND HOW MANY SYSTEMS WERE NEEDED FOR TESTING AND WHERE SHOULD THEY BE PLACED? AS IN MANY HIGH -TECH COMPANIES, THE MICROSOFT TEAM WAS USED TO SHORT, MARKET-DRIVEN PRODUCTION SCHEDULES, BUT THIS TIME THEY WERE IN FOR A SHOCK.

- ORGANIZING THE TESTING TOOK MUCH LONGER THAN THEY ANTICIPATED, BUT THEY SOON LEARNED TO SET REALISTIC EXPECTATIONS THAT WERE IN SYNCH WITH HOSPITAL ACTIVITY AND THE UNEXPECTED DELAYS THAT OCCUR WHEN WORKING WITH PEOPLE WHO ARE UNWELL.

7.3.2 HOW WAS THE TESTING DONE?

- THE TEAM RAN TWO MAIN SETS OF USER TESTS. THE FIRST SET OF TESTS WAS INFORMALLY RUN ONSITE AT THE FRED HUTCHINSON CENTER IN THE HOSPITAL SETTING. AFTER OBSERVING THE SYSTEM IN USE ON COMPUTERS LOCATED IN THE HOSPITAL SETTING, THE TEAM REDESIGNED THE SOFTWARE AND THEN RAN FORMAL USABILITY TESTS IN THE USABILITY LABS AT MICROSOFT.

- **TEST 1 : EARLY OBSERVATIONS ONSITE**

- IN THE INFORMAL TEST AT THE HOSPITAL, SIX COMPUTERS WERE SET UP AND MAINTAINED BY HUTCH STAFF MEMBERS. A SIMPLE, SCALED -BACK PROTOTYPE OF HUTCHWORLD WAS BUILT USING THE EXISTING PRODUCT, MICROSOFT V-CHAT AND WAS INSTALLED ON THE COMPUTERS,

- WHICH PATIENTS AND THEIR FAMILIES FROM VARIOUS HOSPITAL LOCATIONS USED. OVER THE COURSE OF SEVERAL MONTHS, THE TEAM TRAINED HUTCH VOLUNTEERS AND HOSTED EVENTS IN

- THE V-CHAT PROTOTYPE. THE TEAM OBSERVED THE USAGE OF THE SPACE DURING UNSCHEDULED TIMES, AND THEY ALSO OBSERVED THE GENERAL USAGE OF THE PROTOTYPE.

- **TEST 1 : WHAT WAS LEARNED?**

- THIS V-CHAT TEST BROUGHT UP MAJOR USABILITY ISSUES. FIRST, THE USER COMMUNITY WAS RELATIVELY SMALL, AND THERE WERE NEVER ENOUGH PARTICIPANTS IN THE CHAT ROOM FOR SUCCESSFUL COMMUNICATION-A CONCEPT KNOWN AS ***CRITICAL MASS***. IN ADDITION, MANY OF THE PATIENTS WERE NOT INTERESTED IN OR SIMULTANEOUSLY AVAILABLE FOR CHATTING. INSTEAD, THEY PREFERRED ASYNCHRONOUS COMMUNICATION, WHICH DOES NOT REQUIRE AN IMMEDIATE RESPONSE. PATIENTS AND THEIR FAMILIES USED THE COMPUTERS FOR EMAIL, JOURNALS, DISCUSSION LISTS, AND THE BULLETIN BOARDS LARGELY BECAUSE THEY COULD BE USED AT ANY TIME AND DID NOT REQUIRE OTHERS TO BE PRESENT AT THE SAME TIME. THE TEAM LEARNED THAT A STRONG ASYNCHRONOUS BASE WAS ESSENTIAL FOR COMMUNICATION.

- THE TEAM ALSO OBSERVED THAT THE USERS USED THE COMPUTERS TO PLAY GAMES AND TO SEARCH THE WEB FOR CANCER SITES APPROVED BY HUTCH CLINICIANS. THIS INFORMATION WAS NOT INCLUDED IN THE VIRTUAL ENVIRONMENT, AND SO USERS WERE FORCED TO USE MANY DIFFERENT APPLICATIONS. **A** MORE "UNIFIED" PLACE TO FIND ALL OF THE HUTCH CONTENT WAS DESIRED THAT LET USERS RAPIDLY SWAP AMONG A VARIETY OF COMMUNICATION, INFORMATION, AND ENTERTAINMENT TASKS.

- **TEST 1 : THE REDESIGN**

- BASED ON THIS TRIAL, THE TEAM REDESIGNED THE SOFTWARE TO SUPPORT MORE ASYNCHRONOUS COMMUNICATION AND TO INCLUDE A VARIETY OF COMMUNICATION, INFORMATION, AND ENTERTAINMENT AREAS. THEY DID THIS BY MAKING HUTCHWORLD FUNCTION AS A PORTAL THAT PROVIDES ACCESS TO INFORMATION -RETRIEVAL TOOLS, COMMUNICATION TOOLS, GAMES, AND OTHER TYPES OF ENTERTAINMENT. OTHER FEATURES WERE INCORPORATED TOO, INCLUDING EMAIL, A BULLETIN BOARD, A TEXT-CHAT, A WEB PAGE CREATION TOOL, AND A WAY OF CHECKING TO SEE IF ANYONE IS AROUND TO CHAT WITH IN THE 3D WORLD.

TEST 2: USABILITY TESTS

- AFTER REDESIGNING THE SOFTWARE, THE TEAM THEN RAN USABILITY TESTS IN THE MICROSOFT USABILITY LABS. SEVEN PARTICIPANTS (FOUR MALE AND THREE FEMALE) WERE TESTED. FOUR
- OF THESE PARTICIPANTS HAD USED CHAT ROOMS BEFORE AND THREE WERE REGULAR USERS. ALL HAD BROWSED THE WEB AND SOME USED OTHER COMMUNICATIONS SOFTWARE. THE PARTICIPANTS WERE TOLD THAT THEY WOULD USE A PROGRAM CALLED HUTCHWORLD THAT WAS DESIGNED TO PROVIDE SUPPORT FOR PATIENTS AND THEIR FAMILIES. THEY WERE THEN GIVEN FIVE MINUTES TO EXPLORE HUTCHWORLD. THEY WORKED INDEPENDENTLY AND WHILE THEY EXPLORED THEY PROVIDED A RUNNING COMMENTARY ON WHAT THEY WERE LOOKING AT, WHAT THEY WERE THINKING, AND WHAT THEY FOUND CONFUSING. THIS COMMENTARY WAS RECORDED ON VIDEO AND SO WERE THE SCREENS THAT THEY VISITED, SO THAT THE MICROSOFT EVALUATOR, WHO WATCHED THROUGH A ONE -WAY MIRROR, HAD A RECORD OF WHAT HAPPENED FOR LATER ANALYSIS. PARTICIPANTS AND THE EVALUATOR INTERACTED VIA A MICROPHONE AND SPEAKERS. WHEN THE FIVE-MINUTE EXPLORATION PERIOD ENDED, THE PARTICIPANTS WERE ASKED TO COMPLETE A SERIES OF *STRUCTURED TASKS* THAT WERE DESIGNED TO TEST PARTICULAR FEATURES OF THE HUTCHWORLD INTERFACE.
- THESE TASKS FOCUSED ON HOW PARTICIPANTS DEALT WITH THEIR VIRTUAL IDENTITY; THAT IS, HOW THEY REPRESENTED THEMSELVES AND WERE PERCEIVED BY OTHERS COMMUNICATED WITH OTHERS GOT THE INFORMATION THEY WANTED FOUND ENTERTAINMENT

7.3.3 WAS IT TESTED AGAIN?

- FOLLOWING THE USABILITY TESTING, THERE WERE MORE ROUNDS OF OBSERVATION AND TESTING WITH SIX NEW PARTICIPANTS, TWO MALES AND FOUR FEMALES. THESE TESTS FOLLOWED THE SAME GENERAL FORMAT AS THOSE JUST DESCRIBED BUT THIS TIME THEY TESTED MULTIPLE USERS AT ONCE, TO ENSURE THAT THE VIRTUAL WORLD SUPPORTED MULTIUSER INTERACTIONS. THE TESTS WERE ALSO MORE DETAILED AND FOCUSED. THIS TIME THE RESULTS WERE MORE POSITIVE, BUT OF COURSE THERE WERE STILL USABILITY PROBLEMS TO BE FIXED. THEN THE QUESTION AROSE: WHAT TO DO NEXT? IN PARTICULAR, HAD THEY DONE ENOUGH TESTING (SEE DILEMMA)?
- AFTER MAKING A FEW MORE FIXES, THE TEAM STOPPED USABILITY TESTING WITH SPECIFIC TASKS. BUT THE STORY DIDN'T END HERE. THE NEXT STEP WAS TO SHOW HUTCHWORLD TO CANCER PATIENTS AND CAREGIVERS IN A FOCUS-GROUP SETTING AT THE FRED HUTCHINSON CANCER RESEARCH CENTER TO GET THEIR FEEDBACK ON THE FINAL VERSION. ONCE THE TEAM MADE ADJUSTMENTS TO HUTCHWORLD IN RESPONSE TO THE FOCUS-GROUP FEEDBACK, THE FINAL STEP WAS TO SEE HOW WELL HUTCHWORLD WORKED IN A REAL CLINICAL ENVIRONMENT. IT WAS THEREFORE TAKEN TO A RESIDENTIAL BUILDING USED FOR LONG TERM PATIENT AND FAMILY STAYS THAT WAS FULLY WIRED FOR INTERNET ACCESS. HERE, THE TEAM OBSERVED WHAT HAPPENED WHEN IT WAS USED IN THIS NATURAL SETTING. IN PARTICULAR, THEY WANTED TO FIND OUT HOW HUTCHWORLD WOULD INTEGRATE WITH OTHER ASPECTS OF PATIENTS' LIVES, PARTICULARLY WITH THEIR MEDICAL CARE ROUTINES AND THEIR ACCESS TO SOCIAL SUPPORT. THIS INFORMAL OBSERVATION ALLOWED THEM TO EXAMINE PATTERNS OF USE AND TO SEE WHO USED WHICH PARTS OF THE SYSTEM, WHEN, AND WHY.

7.3.4 LOOKING *TO* THE FUTURE

- FUTURE STUDIES WERE PLANNED TO EVALUATE THE EFFECTS OF THE COMPUTERS AND THE SOFTWARE IN THE FRED HUTCHINSON CENTER. THE FOCUS OF THESE STUDIES WILL BE THE SOCIAL SUPPORT AND WELLBEING OF PATIENTS AND THEIR CAREGIVERS IN TWO DIFFERENT CONDITIONS. THERE WILL BE A CONTROL CONDITION IN WHICH USERS (I.E., PATIENTS) LIVE IN THE RESIDENTIAL BUILDING WITHOUT COMPUTERS AND AN EXPERIMENTAL CONDITION IN WHICH USERS LIVE IN SIMILAR CONDITIONS BUT WITH COMPUTERS, INTERNET ACCESS, AND HUTCHWORLD. THE TEAM WILL EVALUATE THE USER DATA (PERFORMANCE AND OBSERVATION) AND SURVEYS COLLECTED IN THE STUDY TO INVESTIGATE KEY QUESTIONS, INCLUDING:

1) HOW DOES THE COMPUTER AND SOFTWARE IMPACT THE SOCIAL WELLBEING OF PATIENTS AND THEIR CAREGIVERS?
2) WHAT TYPE OF COMPUTER-BASED COMMUNICATION BEST SUPPORTS THIS PATIENT COMMUNITY?
3) WHAT ARE THE GENERAL USAGE PATTERNS? I.E., WHICH FEATURES WERE USED AND AT
4) WHAT TIME OF DAY WERE THEY USED, ETC.?
5) HOW MIGHT ANY MEDICAL FACILITY USE COMPUTERS AND SOFTWARE LIKE HUTCH-WORLD TO PROVIDE SOCIAL SUPPORT FOR ITS PATIENTS AND CAREGIVERS?

7.4 EVALUATION PARADIGMS AND TECHNIQUES

- BEFORE WE DESCRIBE THE TECHNIQUES USED IN EVALUATION STUDIES, WE SHALL START BY PROPOSING SOME KEY TERMS. WE START WITH THE MUCH-USED TERM USER STUDIES, DEFINED BY ABIGAIL SELLEN AS FOLLOWS: "USER STUDIES ESSENTIALLY INVOLVE LOOKING AT HOW PEOPLE BEHAVE EITHER IN THEIR NATURAL ENVIRONMENTS, OR IN THE LABORATORY, BOTH WITH OLD TECHNOLOGIES AND WITH NEW ONES." ANY KIND OF EVALUATION, WHETHER IT IS A USER STUDY OR NOT, IS GUIDED EITHER EXPLICITLY OR IMPLICITLY BY A SET OF BELIEFS THAT MAY ALSO BE UNDERPINNED BY THEORY. THESE BELIEFS AND THE PRACTICES (I.E., THE METHODS OR TECHNIQUES) ASSOCIATED WITH THEM ARE KNOWN AS AN EVALUATION PARADIGM, WHICH YOU SHOULD NOT CONFUSE WITH THE "INTERACTION PARADIGMS" DISCUSSED IN CHAPTER 2. OFTEN EVALUATION PARADIGMS ARE RELATED TO A PARTICULAR DISCIPLINE IN THAT THEY STRONGLY INFLUENCE HOW PEOPLE FROM THE DISCIPLINE THINK ABOUT EVALUATION.
- EACH PARADIGM HAS PARTICULAR METHODS AND TECHNIQUES ASSOCIATED WITH IT. WE TEND TO TALK ABOUT TECHNIQUES, BUT YOU MAY FIND THAT OTHER BOOKS CALL THEM METHODS. AN EXAMPLE OF THE RELATIONSHIP BETWEEN A PARADIGM AND THE TECHNIQUES USED BY EVALUATORS FOLLOWING THAT PARADIGM CAN BE SEEN FOR USABILITY TESTING, WHICH IS AN APPLIED SCIENCE AND ENGINEERING PARADIGM. THE TECHNIQUES ASSOCIATED WITH USABILITY TESTING ARE: USER TESTING IN A CONTROLLED ENVIRONMENT; OBSERVATION OF USER ACTIVITY IN THE CONTROLLED ENVIRONMENT AND THE FIELD; AND QUESTIONNAIRES AND INTERVIEWS.

7.4.1 EVALUATION PARADIGMS

- IN THIS PART WE IDENTIFY FOUR CORE EVALUATION PARADIGMS:

1)) A "QUICK AND DIRTY" EVALUATION IS A COMMON PRACTICE IN WHICH DESIGNERS INFORMALLY GET FEEDBACK FROM USERS OR CONSULTANTS TO CONFIRM THAT THEIR IDEAS ARE IN LINE WITH USERS' NEEDS AND ARE LIKED. "QUICK AND DIRTY" EVALUATIONS CAN BE DONE AT ANY STAGE AND THE EMPHASIS IS ON FAST INPUT RATHER THAN CAREFULLY DOCUMENTED FINDINGS. FOR EXAMPLE, EARLY IN DESIGN DEVELOPERS MAY MEET INFORMALLY WITH USERS TO GET FEEDBACK ON IDEAS FOR A NEW PRODUCT (HUGHES ET AL., 1994). AT LATER STAGES SIMILAR MEETINGS MAY OCCUR TO TRY OUT AN IDEA FOR AN ICON, CHECK WHETHER A GRAPHIC IS LIKED, OR CONFIRM THAT INFORMATION HAS BEEN APPROPRIATELY CATEGORIZED ON A WEBPAGE. THIS APPROACH IS OFTEN CALLED "QUICK AND DIRTY" BECAUSE IT IS MEANT TO BE DONE IN A SHORT SPACE OF TIME. GETTING THIS KIND OF FEEDBACK IS AN ESSENTIAL INGREDIENT OF SUCCESSFUL DESIGN.

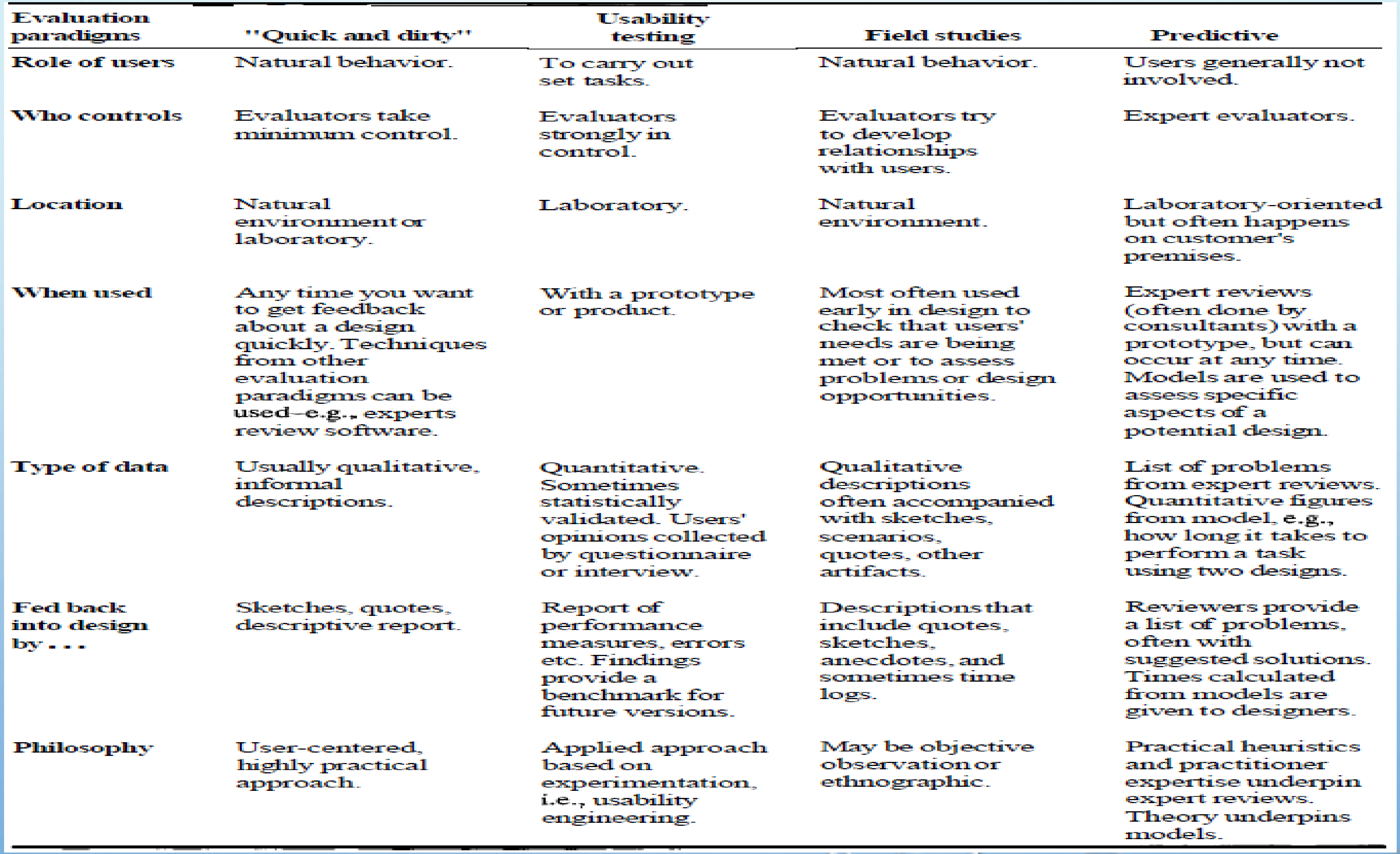
2)) USABILITY TESTING: WAS THE DOMINANT APPROACH IN THE 1980S, AND REMAINS IMPORTANT, ALTHOUGH, AS YOU WILL SEE, FIELD STUDIES AND HEURISTIC EVALUATIONS HAVE GROWN IN PROMINENCE. USABILITY TESTING INVOLVES MEASURING TYPICAL USERS' PERFORMANCE ON CAREFULLY PREPARED TASKS THAT ARE TYPICAL OF THOSE FOR WHICH THE SYSTEM WAS DESIGNED. USERS' PERFORMANCE IS GENERALLY MEASURED IN TERMS OF NUMBER OF ERRORS AND TIME TO COMPLETE THE TASK. AS THE USERS PERFORM THESE TASKS, THEY ARE WATCHED AND RECORDED ON VIDEO AND BY LOGGING THEIR INTERACTIONS WITH SOFTWARE. THIS OBSERVATIONAL DATA IS USED TO CALCULATE PERFORMANCE TIMES, IDENTIFY ERRORS, AND HELP EXPLAIN WHY THE USERS DID WHAT THEY DID. USER SATISFACTION QUESTIONNAIRES AND INTERVIEWS ARE ALSO USED TO ELICIT USERS' OPINIONS.

3)) FIELD STUDIES: THE DISTINGUISHING FEATURE OF FIELD STUDIES IS THAT THEY ARE DONE IN NATURAL SETTINGS WITH THE AIM OF INCREASING UNDERSTANDING ABOUT WHAT USERS DO NATURALLY AND HOW TECHNOLOGY IMPACTS THEM. IN PRODUCT DESIGN, FIELD STUDIES CAN BE USED TO :

- HELP IDENTIFY OPPORTUNITIES FOR NEW TECHNOLOGY
- DETERMINE REQUIREMENTS FOR DESIGN
- FACILITATE THE INTRODUCTION OF TECHNOLOGY
- EVALUATE TECHNOLOGY.

4)) PREDICTIVE EVALUATION: IN PREDICTIVE EVALUATIONS EXPERTS APPLY THEIR KNOWLEDGE OF TYPICAL USERS, OFTEN GUIDED BY HEURISTICS, TO PREDICT USABILITY PROBLEMS. ANOTHER APPROACH INVOLVES THEORETICALLY BASED MODELS. THE KEY FEATURE OF PREDICTIVE EVALUATION IS THAT USERS NEED NOT BE PRESENT, WHICH MAKES THE PROCESS QUICK, RELATIVELY INEXPENSIVE, AND THUS ATTRACTIVE TO COMPANIES; BUT IT HAS LIMITATIONS.

TABLE 7.1 CHARACTERISTICS OF DIFFERENT EVALUATION PARADIGMS



Evaluation paradigms	"Quick and dirty"	Usability testing	Field studies	Predictive
Role of users	Natural behavior.	To carry out set tasks.	Natural behavior.	Users generally not involved.
Who controls	Evaluators take minimum control.	Evaluators strongly in control.	Evaluators try to develop relationships with users.	Expert evaluators.
Location	Natural environment or laboratory.	Laboratory.	Natural environment.	Laboratory-oriented but often happens on customer's premises.
When used	Any time you want to get feedback about a design quickly. Techniques from other evaluation paradigms can be used–e.g., experts review software.	With a prototype or product.	Most often used early in design to check that users' needs are being met or to assess problems or design opportunities.	Expert reviews (often done by consultants) with a prototype, but can occur at any time. Models are used to assess specific aspects of a potential design.
Type of data	Usually qualitative, informal descriptions.	Quantitative. Sometimes statistically validated. Users' opinions collected by questionnaire or interview.	Qualitative descriptions often accompanied with sketches, scenarios, quotes, other artifacts.	List of problems from expert reviews. Quantitative figures from model, e.g., how long it takes to perform a task using two designs.
Fed back into design by . . .	Sketches, quotes, descriptive report.	Report of performance measures, errors etc. Findings provide a benchmark for future versions.	Descriptions that include quotes, sketches, anecdotes, and sometimes time logs.	Reviewers provide a list of problems, often with suggested solutions. Times calculated from models are given to designers.
Philosophy	User-centered, highly practical approach.	Applied approach based on experimentation, i.e., usability engineering.	May be objective observation or ethnographic.	Practical heuristics and practitioner expertise underpin expert reviews. Theory underpins models.

7.4.2 TECHNIQUES

THERE ARE MANY EVALUATION TECHNIQUES AND THEY CAN BE CATEGORIZED IN VARIOUS WAYS, BUT IN THIS TEXT WE WILL EXAMINE TECHNIQUES FOR:

- OBSERVING USERS
- ASKING USERS THEIR OPINIONS
- ASKING EXPERTS THEIR OPINIONS
- TESTING USERS' PERFORMANCE
- MODELING USERS' TASK PERFORMANCE TO PREDICT THE EFFICACY OF A USER INTERFACE

THE BRIEF DESCRIPTIONS BELOW OFFER AN OVERVIEW OF EACH CATEGORY, BE AWARE THAT SOME TECHNIQUES ARE USED IN DIFFERENT WAYS IN DIFFERENT EVALUATION PARADIGMS.

1. **OBSERVING USERS:** OBSERVATION TECHNIQUES HELP TO IDENTIFY NEEDS LEADING TO NEW TYPES OF PRODUCTS AND HELP TO EVALUATE PROTOTYPES. NOTES, AUDIO, VIDEO, AND INTERACTION LOGS ARE WELL KNOWN WAYS OF RECORDING OBSERVATIONS AND EACH HAS BENEFITS AND DRAWBACKS. OBVIOUS CHALLENGES FOR EVALUATORS ARE HOW TO OBSERVE WITHOUT DISTURBING THE PEOPLE BEING OBSERVED AND HOW TO ANALYZE THE DATA, PARTICULARLY WHEN LARGE QUANTITIES OF VIDEO DATA ARE COLLECTED OR WHEN SEVERAL DIFFERENT TYPES MUST BE INTEGRATED TO TELL THE STORY (E.G., NOTES, PICTURES, AND SKETCHES FROM OBSERVERS).

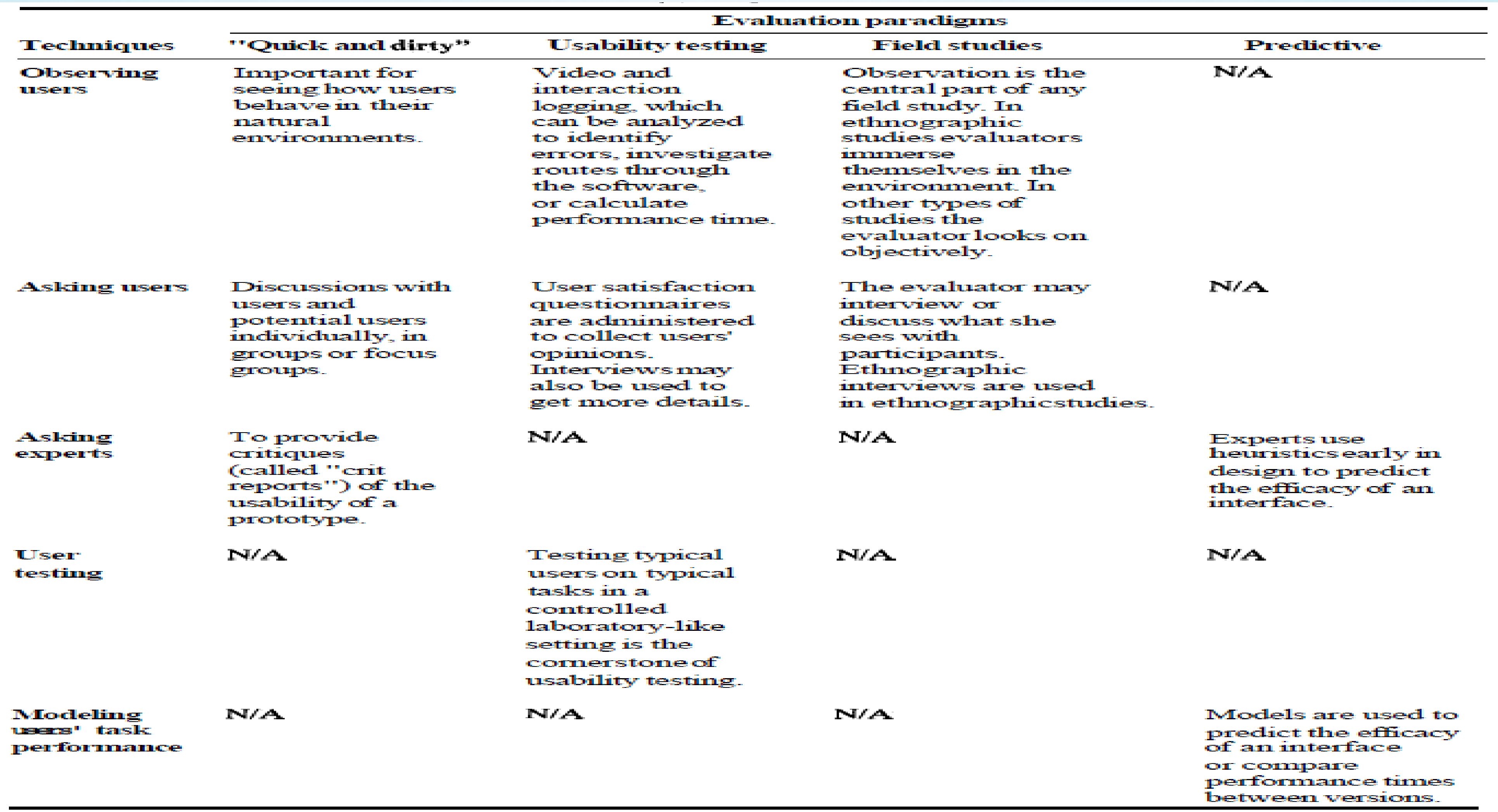
2. **ASKING USERS:** WHAT THEY THINK OF A PRODUCT-WHETHER IT DOES WHAT THEY WANT; WHETHER THEY LIKE IT; WHETHER THE AESTHETIC DESIGN APPEALS; WHETHER THEY HAD PROBLEMS USING IT; WHETHER THEY WANT TO USE IT AGAIN-IS AN OBVIOUS WAY OF GETTING FEEDBACK. INTERVIEWS AND QUESTIONNAIRES ARE THE MAIN TECHNIQUES FOR DOING THIS. THE QUESTIONS ASKED CAN BE UNSTRUCTURED OR TIGHTLY STRUCTURED. THEY CAN BE ASKED OF A FEW PEOPLE OR OF HUNDREDS. INTERVIEW AND QUESTIONNAIRE TECHNIQUES ARE ALSO BEING DEVELOPED FOR USE WITH EMAIL AND THE WEB.

3. **ASKING EXPERTS:** SOFTWARE INSPECTIONS AND REVIEWS ARE LONG ESTABLISHED TECHNIQUES FOR EVALUATING SOFTWARE CODE AND STRUCTURE. DURING THE 1980S VERSIONS OF SIMILAR TECHNIQUES WERE DEVELOPED FOR EVALUATING USABILITY. GUIDED BY HEURISTICS, EXPERTS STEP THROUGH TASKS ROLE-PLAYING TYPICAL USERS AND IDENTIFY PROBLEMS. DEVELOPERS LIKE THIS APPROACH BECAUSE IT IS USUALLY RELATIVELY INEXPENSIVE AND QUICK TO PERFORM COMPARED WITH LABORATORY AND FIELD EVALUATIONS THAT INVOLVE USERS. IN ADDITION, EXPERTS FREQUENTLY SUGGEST SOLUTIONS TO PROBLEMS.

4. **USER TESTING:** MEASURING USER PERFORMANCE TO COMPARE TWO OR MORE DESIGNS HAS BEEN THE BEDROCK OF USABILITY TESTING. AS WE SAID EARLIER WHEN DISCUSSING USABILITY TESTING, THESE TESTS ARE USUALLY CONDUCTED IN CONTROLLED SETTINGS AND INVOLVE TYPICAL USERS PERFORMING TYPICAL, WELL-DEFINED TASKS. DATA IS COLLECTED SO THAT PERFORMANCE CAN BE ANALYZED. GENERALLY THE TIME TAKEN TO COMPLETE A TASK, THE NUMBER OF ERRORS MADE, AND THE NAVIGATION PATH THROUGH THE PRODUCT ARE RECORDED. DESCRIPTIVE STATISTICAL MEASURES SUCH AS MEANS AND STANDARD DEVIATIONS ARE COMMONLY USED TO REPORT THE RESULTS.

5. **MODELING USERS' TASK PERFORMANCE:** THERE HAVE BEEN VARIOUS ATTEMPTS TO MODEL HUMAN-COMPUTER INTERACTION SO AS TO PREDICT THE EFFICIENCY AND PROBLEMS ASSOCIATED WITH DIFFERENT DESIGNS AT AN EARLY STAGE WITHOUT BUILDING ELABORATE PROTOTYPES. THESE TECHNIQUES ARE SUCCESSFUL FOR SYSTEMS WITH LIMITED FUNCTIONALITY SUCH AS TELEPHONE SYSTEMS ARE THE BEST KNOWN TECHNIQUES.

TABLE 7.1 SUMMARIZES THE CATEGORIES OF TECHNIQUES AND INDICATES HOW THEY ARE COMMONLY USED IN THE FOUR EVALUATION PARADIGMS.



	Evaluation paradigms			
Techniques	"Quick and dirty"	Usability testing	Field studies	Predictive
Observing users	Important for seeing how users behave in their natural environments.	Video and interaction logging, which can be analyzed to identify errors, investigate routes through the software, or calculate performance time.	Observation is the central part of any field study. In ethnographic studies evaluators immerse themselves in the environment. In other types of studies the evaluator looks on objectively.	N/A
Asking users	Discussions with users and potential users individually, in groups or focus groups.	User satisfaction questionnaires are administered to collect users' opinions. Interviews may also be used to get more details.	The evaluator may interview or discuss what she sees with participants. Ethnographic interviews are used in ethnographicstudies.	N/A
Asking experts	To provide critiques (called "crit reports") of the usability of a prototype.	N/A	N/A	Experts use heuristics early in design to predict the efficacy of an interface.
User testing	N/A	Testing typical users on typical tasks in a controlled laboratory-like setting is the cornerstone of usability testing.	N/A	N/A
Modeling users' task performance	N/A	N/A	N/A	Models are used to predict the efficacy of an interface or compare performance times between versions.

7.5 DECIDES: A FRAMEWORK TO GUIDE EVALUATION

WELL-PLANNED EVALUATIONS ARE DRIVEN BY CLEAR GOALS AND APPROPRIATE QUESTIONS. TO GUIDE OUR EVALUATIONS WE USE THE DECIDE FRAMEWORK, WHICH PROVIDES THE FOLLOWING CHECKLIST TO HELP NOVICE EVALUATORS:

1. DETERMINE THE OVERALL GOALS THAT THE EVALUATION ADDRESSES.

2. EXPLORE THE SPECIFIC QUESTIONS TO BE ANSWERED.

3. CHOOSE THE EVALUATION PARADIGM AND TECHNIQUES TO ANSWER THE QUESTIONS.

4. IDENTIFY THE PRACTICAL ISSUES THAT MUST BE ADDRESSED, SUCH AS SELECTING PARTICIPANTS.

5. DECIDE HOW TO DEAL WITH THE ETHICAL ISSUES.

6. EVALUATE, INTERPRET, AND PRESENT THE DATA.

7.6 DISCUSSION

IN BOTH HUTCHWORLD AND THE 1984 OLYMPIC MESSAGING SYSTEM, A VARIETY OF EVALUATION TECHNIQUES WERE USED AT DIFFERENT STAGES OF DESIGN TO ANSWER DIFFERENT QUESTIONS. "QUICK AND DIRTY" OBSERVATION, IN WHICH THE EVALUATORS INFORMALLY EXAMINE HOW A PROTOTYPE IS USED IN THE NATURAL ENVIRONMENT, WAS VERY USEFUL IN EARLY DESIGN. FOLLOWING THIS WITH ROUNDS OF USABILITY TESTING AND REDESIGN REVEALED IMPORTANT USABILITY PROBLEMS. HOWEVER, USABILITY TESTING ALONE IS NOT SUFFICIENT. FIELD STUDIES WERE NEEDED TO SEE HOW USERS USED THE SYSTEM IN THEIR NATURAL ENVIRONMENTS, AND SOMETIMES THE RESULTS WERE SURPRISING. FOR EXAMPLE, IN THE OMS SYSTEM USERS FROM DIFFERENT CULTURES BEHAVED DIFFERENTLY. A KEY ISSUE IN THE HUTCHWORLD STUDY WAS HOW USE OF THE SYSTEM WOULD FIT WITH PATIENTS' MEDICAL ROUTINES AND CHANGES IN THEIR PHYSICAL AND EMOTIONAL STATES. USERS' OPINIONS ALSO OFFERED VALUABLE INSIGHTS. AFTER ALL, IF USERS DON'T LIKE A SYSTEM, IT DOESN'T MATTER HOW SUCCESSFUL THE USABILITY TESTING IS: THEY PROBABLY WON'T USE IT. QUESTIONNAIRES AND INTERVIEWS WERE USED TO COLLECT USER'S OPINIONS.

- AN INTERESTING POINT CONCERNS NOT ONLY HOW THE DIFFERENT TECHNIQUES CAN BE USED TO ADDRESS DIFFERENT ISSUES AT DIFFERENT STAGES OF DESIGN, BUT ALSO HOW THESE TECHNIQUES COMPLEMENT EACH OTHER. TOGETHER THEY PROVIDE A BROAD PICTURE OF THE SYSTEM'S USABILITY AND REVEAL DIFFERENT PERSPECTIVES. IN ADDITION, SOME TECHNIQUES ARE BETTER THAN OTHERS FOR GETTING AROUND PRACTICAL PROBLEMS. THIS IS A LARGE PART OF BEING A SUCCESSFUL EVALUATOR. IN THE HUTCHWORLD STUDY, FOR EXAMPLE, THERE WERE NOT MANY USERS, SO THE EVALUATORS NEEDED TO INVOLVE THEM SPARINGLY. FOR EXAMPLE, A TECHNIQUE REQUIRING 20 USERS TO BE AVAILABLE AT THE SAME TIME WAS NOT FEASIBLE IN THE HUTCHWORLD STUDY, WHEREAS THERE WAS NO PROBLEM WITH SUCH AN APPROACH IN THE OMS STUDY. FURTHERMORE, THE OMS STUDY ILLUSTRATED HOW MANY DIFFERENT TECHNIQUES, SOME OF WHICH WERE HIGHLY OPPORTUNISTIC, CAN BE BROUGHT INTO PLAY DEPENDING ON CIRCUMSTANCES. SOME PRACTICAL ISSUES THAT EVALUATORS ROUTINELY HAVE TO ADDRESS INCLUDE:
    - WHAT TO DO WHEN THERE ARE NOT MANY USERS
    - HOW TO OBSERVE USERS IN THEIR NATURAL LOCATION (I.E., FIELD STUDIES) WITHOUT DISTURBING THEM
    - HAVING APPROPRIATE EQUIPMENT AVAILABLE
    - DEALING WITH SHORT SCHEDULES AND LOW BUDGETS
    - NOT DISTURBING USERS OR CAUSING THEM DURESS OR DOING ANYTHING UNETHICAL
    - COLLECTING "USEFUL" DATA AND BEING ABLE TO ANALYZE IT
    - SELECTING TECHNIQUES THAT MATCH THE EVALUATORS' EXPERTISE.