L 77

中华人民共和国国家标准

GB/T 16260.4—2006/ISO/IEC TR 9126-4:2004

软件工程　产品质量
第4部分:使用质量的度量

Software engineering—
Product quality—
Part 4:Quality in use metrics

(ISO/IEC TR 9126-4:2004,IDT)

2006-03-14 发布　　2006-07-01 实施

中华人民共和国国家质量监督检验检疫总局
中国国家标准化管理委员会　发布

目　次

前　言

GB/T 16260《软件工程　产品质量》分为如下几部分：

——第1部分(即GB/T 16260.1)：质量模型；

——第2部分(即GB/T 16260.2)：外部度量；

——第3部分(即GB/T 16260.3)：内部度量；

——第4部分(即GB/T 16260.4)：使用质量的度量。

本部分为GB/T 16260的第4部分。

本部分等同采用ISO/IEC TR 9126-4:2004《软件工程　产品质量　第4部分：使用质量的度量》。

为便于使用，本部分做了下列编辑性修改：

a) “本技术报告”改为“本部分”；

b) 删除了国际标准的前言，修改了国际标准的引言；

c) 由于原国际标准ISO 8402:1994已废止，因而本部分第3章删去了ISO/IEC TR 9126-4:2004中第3章对这个文件的引用；同时，据此原因，本部分中D.1.1也删去了原文中的注1，使本部分的附录D与GB/T 16260第2、3部分的附录D保持一致。

本部分的附录A、附录B、附录C、附录D、附录E、附录F和附录G是资料性附录。

本部分由中华人民共和国信息产业部提出。

本部分由中国电子技术标准化研究所归口。

本部分起草单位：上海计算机软件技术开发中心、中国电子技术标准化研究所、浙江省电子产品检验所。

本部分主要起草人：杨根兴、葛孝堃、韩良秀、冯惠、王凌、宣以广。

引　言

GB/T 16260 的本部分所述的使用质量的度量是用来测量 GB/T 16260.1—2006 中所定义的使用质量的属性。本部分所列的度量并非一个完备的度量集合。开发者、评价者、质量管理者及需方可以从中选择度量来定义需求、评价软件产品、测量质量和用于其他目的。他们也可以修改这些度量或使用本部分未包括的其他度量。本部分适用于各种软件产品，但是并非每种度量都适用于每种软件产品。

GB/T 16260.1—2006 定义了软件质量特性，及这些特性又如何分解成各个子特性的相关术语。但在该部分中，并没有描述这些子特性是如何被测量的。对于这些特性或子特性的测量 GB/T 16260.2—2006 定义了外部度量，GB/T 16260.3—2006 定义了内部度量，GB/T 16260.4—2006 定义了使用质量的度量。内部度量用来测量软件本身；外部度量用来测量包括软件在内的基于计算机系统的行为；而使用质量的度量则是测量软件在某个特定使用周境中的使用效果。

本部分旨在和 GB/T 16260.1—2006 结合使用。因此，在使用本部分以前，极力推荐读者先阅读 GB/T 18905.1—2002 和 GB/T 16260.1—2006。尤其是读者不熟悉在产品规格说明和评价中如何使用软件度量的情况下。

软件工程　产品质量
第4部分:使用质量的度量

1　范围

GB/T 16260的本部分为GB/T 16260.1—2006中所规定的质量特性定义了使用质量的度量。本部分旨在与GB/T 16260.1—2006一起使用。

本部分包括以下内容:

a)　如何应用软件质量度量的解释;

b)　每种特性的基本度量集;

c)　在软件产品生存周期中如何应用这些度量的实例。

它包括了作为附录的使用质量的评价过程和报告格式。

本部分没有为某个评定级别或依从性等级设置这些度量值的范围,因为这些值是依据每个软件产品或软件产品的一部分的自身特性而定的,即依赖于软件分类、完整性级别和用户需求等因素。一些属性可能会有期望的取值范围,但不依赖于特定用户的需求,而范围的确定往往依赖于一般因素,例如人类认知因素。

本部分可用于各种应用软件。用户可以选择、修改及应用本部分中的度量和测度,也可以针对独特应用领域定义特定的应用的度量。例如,对于安全性和安全保密性等质量特性的具体测量可参见有关国家标准和国际标准。

本部分旨在针对以下使用者:

a)　需方(从供方获得或采购系统、软件产品或软件服务的个体或组织);

b)　评价者(实施评价的个体或组织。例如评价者可以是测试实验室、软件开发组织的质量部门、政府组织或用户);

c)　开发者(执行开发活动的个体或组织,开发活动包括软件生存周期过程中的需求分析、设计、测试直至验收等活动);

d)　维护者(执行维护活动的个体或组织);

e)　供方(按所签合同向需方提供系统、软件产品或软件服务的个体或组织),其在合格性测试中确认软件质量时使用;

f)　用户(使用软件产品执行具体功能的个体或组织),其在验收测试中评价软件产品质量时使用;

g)　质量管理者(执行软件产品或软件服务的系统性检查的个体或组织),其在质量保证和质量控制中评价软件质量时使用。

2　符合性

符合性不作要求。

注:在GB/T 16260.1—2006质量模型中有关于度量的一般符合性要求。

3　规范性引用文件

下列文件中的条款通过GB/T 16260的本部分的引用而成为本部分的条款。凡是注日期的引用文件,其随后所有的修改单(不包括勘误的内容)或修订版均不适用于本部分,然而,鼓励根据本部分达成协议的各方研究是否可使用这些文件的最新版本。凡是不注日期的引用文件,其最新版本适用于本

部分。

GB/T 8566—2001 信息技术 软件生存周期过程(idt ISO/IEC 12207:1995)

GB/T 16260.1—2006 软件工程 产品质量 第1部分:质量模型(ISO/IEC 9126-1:2001,IDT)

GB/T 16260.2—2006 软件工程 产品质量 第2部分:外部度量(ISO/IEC TR 9126-2:2003,IDT)

GB/T 16260.3—2006 软件工程 产品质量 第3部分:内部度量(ISO/IEC TR 9126-3:2003,IDT)

GB/T 18905.1—2002 软件工程 产品评价 第1部分:概述(ISO/IEC 14598-1:1999,IDT)

GB/T 18905.2—2002 软件工程 产品评价 第2部分:策划和管理(ISO/IEC 14598-2:2000,IDT)

GB/T 18905.3—2002 软件工程 产品评价 第3部分:开发者用的过程(ISO/IEC 14598-3:2000,IDT)

GB/T 18905.4—2002 软件工程 产品评价 第4部分:需方用的过程(ISO/IEC 14598-4:1999,IDT)

GB/T 18905.5—2002 软件工程 产品评价 第5部分:评价者用的过程(ISO/IEC 14598-5:1998,IDT)

GB/T 18905.6—2002 软件工程 产品评价 第6部分:评价模块的文档编制(ISO/IEC 14598-6:2001,IDT)

GB/T 18491.1—2001 信息技术 软件测量 功能规模测量 第1部分:概念定义(idt ISO/IEC 14143-1:1998)

ISO 9241-11:1998 使用视觉显示终端(VDTs)办公的人类工效学要求 易用性指南

4 术语和定义

在GB/T 18905.1—2002和GB/T 16260.1—2006中定义的术语适用于本部分,并且这些术语列在了附录D中。

4.1

使用周境 context of use

用户、任务、设备(硬件、软件和材料)及产品使用的物理和社会环境。

[ISO 9241-11:1998]

4.2

目标 goal

预期的结果。

[ISO 9241-11:1998]

4.3

任务 task

实现目标所必需的活动。

注1:这些活动可能是体力的或脑力的。

注2:工作职责可以决定目标和任务。

[ISO 9241-11:1998]

5 符号和缩略语

SQA 软件质量保证(组)

SLCP 软件生存周期过程

6 软件质量度量的使用

GB/T 16260 的第 2、第 3 和第 4 部分提出了与第 1 部分“质量模型”一起使用的一组质量度量(外部质量、内部质量和使用质量的度量)的建议。这些部分的用户可以修改已定义的度量,也可以使用未列出的度量。当使用一个已修改的或一个未在各部分中定义的新度量时,用户宜说明这些度量与第 1 部分中的质量模型或任何其他所用的替代质量模型之间的关系。

GB/T 16260 的用户宜从第 1 部分中选择用于评价的质量特性和子特性,确定要采用的适当的直接测度和间接测度,确定相关的度量,并以客观的方式解释测量结果。GB/T 16260 的用户也可以从 GB/T 18905 系列标准中选择软件生存周期中的产品质量评价过程。上述这些标准给出了测量、评估和评价软件产品质量的方法,旨在供开发者、需方和独立的评价者使用,特别是那些负责软件产品评价的人员(见图 1)。

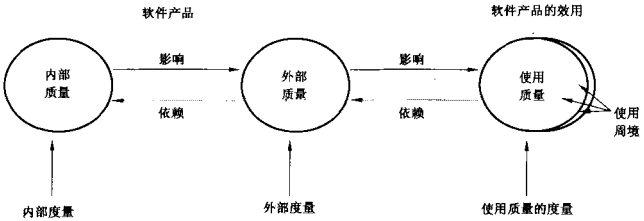


图 1 不同度量类型之间的关系

内部度量可用于开发阶段的非执行软件产品(例如标书、需求定义、设计规格说明或源代码等)。内部度量为用户提供了测量中间可交付项的质量的能力,从而可以预计最终产品的质量。这样就可使用户尽可能在软件生存周期的开发阶段的早期识别质量问题,并实施纠正措施。

外部度量可以通过测量该软件产品作为其一部分的系统的行为来测量软件产品的质量。外部度量只能在生存周期过程的测试阶段和任何运行阶段使用。当在其所在系统环境下运行软件产品时就可以执行这样的测量。

使用质量的度量是测量产品在特定的使用周境中,是否满足特定用户达到特定目标所要求的有效性、生产率、安全性及满意度。这只能在真实的系统环境中获得。

用户的质量要求可用使用质量的度量、外部度量,甚至是内部度量的质量需求来说明。这些由度量规定的需求宜作为产品评价时的准则。

建议尽可能采用与目标外部度量有密切关系的内部度量,以便可以用这些内部度量来预计外部度量的值。然而,往往很难设计出一个能够在内部和外部度量间提供密切关系的严格的理论模型。因此,假设模型可能是模糊的,所以在使用度量时,外部度量和内部度量关系密切程度模型应该使用统计建模的方法。

GB/T 16260 的第 1 部分附录 A 中的 A.4 列出了与有效性和可信赖性相关的建议和需求。另外,本部分的附录 A 列出了使用度量时的一些考虑细节。

7 度量表的阅读和使用

第 8 章列出的度量依据 GB/T 16260.1—2006 中的特性和子特性进行分类。下面是表中的每个度量应给出的信息:

a) 度量名称

在内部度量表和外部度量表中的相应度量的类似名称。

b) 度量目的

在度量应用中以回答问题的形式进行描述。

c) 应用的方法

提供一个应用的大纲。

d) 测量、公式和数据元素计算

给出测量公式,并解释所用的数据元素的意义。

注:在某些情况下一个度量对应多个公式。

e) 测量值解释

给出范围和最佳值。

f) 度量标度类型

度量中使用的标度类型。包括:标称标度、顺序标度、间隔标度、比率标度和绝对标度。

注:附录C有详细的解释。

g) 测度类型

所用的类型是:规模类型(例如功能规模、源代码规模)、时间类型(例如经时时间、用户时间)、计数类型(变更数、失效数)。

注:附录C有详细的解释。

h) 测量输入

测量中使用的数据来源。

i) 在GB/T 8566中的应用

标识出应用度量的软件生存周期过程。

j) 目标用户

标识测量结果的使用者。

8 度量表

8.0 综述

在本章列出的所有度量不是一个完备的集合,而且可能还有未经确认的。这些度量按照GB/T 16260.1—2006中说明的软件质量特性的顺序列出。

可应用的度量并不局限于这里所列的。其他特定用途的具体度量将由其他相关文档给出,如功能规模测量或精确时间的效率测量。

注:推荐选用有关标准、技术报告或指南中规定的度量或测量。功能规模的测量定义见GB/T 18491.1—2001。精确时间的效率测量的实例见ISO/IEC 14756。

在一个特定环境中应用度量前,应确认该度量(见附录A)。

本章所列的使用质量的度量是测量产品在特定的使用周境中,满足特定用户达到特定目标所要求的有效性、生产率、安全性及满意度的程度。使用质量不仅依赖于软件产品,而且依赖于产品使用的特定周境,而使用周境是由用户、任务、物理或社会环境因素决定的。

在实际的使用周境中,通过观察典型用户完成的典型任务来评估使用质量(见附录E)。这些测度可以通过模拟一个实际的工作环境(如易用性实验室)或是观察产品的运行使用情况来获得。为了说明或测量使用质量,首先必须要明确使用周境中的各个部分:用户、用户目标及使用环境。评价的设计应该和这些使用周境尽可能地相匹配。不过在运行环境下为用户提供有利的帮助和协助的方式也是重要的。

注:在某些情况下,易用性和使用质量有相类似的意义(安全性除外)(如ISO 9241-11)。

一些外部的易用性度量(GB/T 16260.2—2006)使用类似的方法进行测试。但对特定产品特征使用的评价,要在产品更为通用的使用期内完成具有代表性任务(作为测试使用质量一部分)时进行。

使用质量有四个特性:有效性、生产率、安全性和满意度,没有子特性。

8.1 有效性度量

有效性度量(见表 8.1)评估的是在特定的使用周境中,用户执行任务时是否能够准确和完全地达到规定的目标。这种度量只考虑已经完成目标的程度,而不考虑是如何达到目标的。

8.2 生产率度量

生产率度量(见表 8.2)评估的是在特定的使用周境中用户消耗的与所达到的有效性相关的资源。虽然其他相关资源可能包含用户的工作量、材料或使用的财务成本,但最常见的资源是完成任务的时间。

8.3 安全性度量

安全性度量(见表 8.3)评估的是在特定的使用周境中对人、业务、软件、财产或环境产生伤害的风险级别。包括用户以及那些受使用影响的人的健康和安全以及意想不到的生理的或经济的后果。

8.4 满意度度量

满意度度量(见表 8.4)评估的是在特定的使用周境中用户对产品使用的态度。

注:满意度受用户对软件产品的属性的感知(例如可由那些外部度量所测得的)和用户对使用中的软件效率、生产率以及安全性的感知的影响。

表 8.1 有效性度量

度量名称	度量目的	应用的方法	测量、公式及数据元素计算	测量值解释	度量标度类型	测度类型	测量输入	在 GB/T8566 中的应用	目标用户
任务有效性	已正确完成任务目标的比例是多少?	用户测试	$M_1=\mid 1-\Sigma A_i \mid$ A_i=任务输出中遗漏或不正确的组件的比例值	$0.0\leqslant M_1\leqslant 1.0$ 越接近 1.0 越好		A=比例	运行(测试)报告 用户监控记录 用户监视记录	6.5 确认 5.3.9 合格性测试 5.4 运作	用户 人机界面设计者
注:根据每项潜在遗漏的或不完全的任务输出部件给业务或用户带来的损失程度,分别定出相应的权值 A_i(如果权值总和大于 1,度量通常得 0,即使这可能暗示着负的结果和潜在的安全隐患。)(见例 G 3.1.1)记分机制是通过将其应用于一系列的任务输出,并调整权值直至测度具有可重复性、可再现性且有意义后经迭代确定出来的。									
任务完成量	已完成的任务比例是多少?	用户测试	$X=A/B$ A=已完成的任务数 B=试图完成的总任务数	$0.0\leqslant X\leqslant 1.0$ 越接近 1.0 越好	比率标度	A=计数 B=计数 X = 计数/计数	运行(测试)报告 用户监控记录	6.5 确认 5.3.9 合格性测试 5.4 运作	用户 人机界面设计者
注:度量可以通过测量一个或一组用户得出。如果任务可以部分地完成,那么就可以使用任务有效性度量了。									
出错频率	出错频率怎样?	用户测试	$X=A/T$ A=用户导致的错误数 T=任务时间或任务数。	$0\leqslant X$ 越接近 0 越好	绝对标度	A=计数	运行(测试)报告 用户监控记录	6.5 确认 5.3.9 合格性测试 5.4 运作	用户 人机界面设计者
注:这种度量仅适用于错误具有等同权重或定义了权值的情况下做出的比较。									

表 8.2 生产率度量

度量名称	度量目的	应用的方法	测量、公式及数据元素计算	测量值解释	度量标度类型	测度类型	测量输入	在 GB/T 8566 中的应用	目标用户
任务时间	完成一项任务所需的时间?	用户测试	$X=T_a$ T_a=任务时间	$0.0 \leqslant X$ 越小越好	间隔标度	T=时间	运行(测试)报告 用户监视记录	6.5 确认 5.3.9 合格性测试 5.4 运作	用户 人机界面设计者
任务效率	用户的效率如何?	用户测试	$X=M_t/T$ M_t=任务的有效时间 T=任务时间	$0.0 \leqslant X$ 越大越好		T=时间 X=比例/时间	运行(测试)报告 用户监视记录	6.5 确认 5.3.9 合格性测试 5.4 运作	用户 人机界面设计者
注 1:任务效率测量每个单位时间内所完成目标的比率。高比值表明在短期内任务的完成比例高。从而任务效率具有了可比性,如将速度快但容易出错的界面和速度慢但操作简单的界面的比较(见 F.2.4.4)。 注 2:如果已经测量出了任务完成量,那么任务效率可以用任务完成量/任务时间来得到。它可测量出每单位时间成功用户的比例。高比值表明在短时间内就有高比例成功的用户。									
经济生产率	用户的成本有效性如何?	用户测试	$X=M_t/C$ M_t=任务有效成本 C=任务总成本	$0.0 \leqslant X$ 越大越好		T=时间 X=	运行(测试)报告 用户监视记录	6.5 确认 5.3.9 合格性测试 5.4 运作	用户 人机界面设计者
注:成本包括如用户花费的时间,其他人提供帮助的时间,计算资源、电话、材料的开销。									
生产比例	用户实施生产行为的时间比率怎样?	用户测试	$X=T_a/T_b$ T_a=生产时间=任务时间−帮助时间−出错时间−搜索时间 T_b=任务时间	$0.0 \leqslant X \leqslant 1.0$ 越趋近于 1.0 越好	绝对标度	T_a=时间 T_b=时间 X=时间/时间	运行(测试)报告 用户监视记录	6.5 确认 5.3.9 合格性测试 5.4 运作	用户 人机界面设计者
注:这种度量要求对具有相互作用的录影带进行详尽的分析。(见 Macleod M,Bowden R,Bevan N 和 Curson I (1997)The MUSIC Performance Measurement method,Behaviour and information Technology,16,279—293。)									
相对的用户效率	与专家相比一个普通用户的效率如何?	用户测试	相对的用户效率 $X=A/B$ A=普通用户的任务效率 B=专家用户的任务效率	$0.0 \leqslant X \leqslant 1.0$ 越趋近于 1.0 越好	绝对标度	X=比例/比例	运行(测试)报告 用户监视记录	6.5 确认 5.3.9 合格性测试 5.4 运作	用户 人机界面设计者
注:普通用户和专家执行同样的任务。假设专家有 100%的生产率,普通用户和专家有同样的有效性,那么这种度量将得到一个与生产比例相近的值。									

表 8.3 安全性度量

度量名称	度量目的	应用的方法	测量、公式及数据元素计算	测量值解释	度量标度类型	测度类型	测量输入	在 GB/T 8566 中的应用	目标用户
用户健康和安全	用户受到健康问题的影响范围怎样?	使用统计	$X=1-A/B$ A=报告有 RSI 的用户数 B=用户总数	$0.0 \leqslant X \leqslant 1.0$ 越趋近于 1.0 越好	绝对标度	A=计数 B=计数 X = 计数/计数	使用监控记录	5.4 运作	用户 人机界面设计者
注:健康问题可以包括重复性的劳损(RSI)、疲倦、头痛等。									
使用该系统对人身安全的影响	用户使用系统所遇到灾难的影响范围怎样?	使用统计	$X=1-A/B$ A=遇到灾难的用户数 B=受系统影响的潜在用户总数	$0.0 \leqslant X \leqslant 1.0$ 越趋近于 1.0 越好	绝对标度	A=计数 B=计数 X = 计数/计数	使用监控记录	5.3.9 合格性测试 5.4 运作	用户 人机界面设计者 开发者
注:患者的安全性是这种度量的一个实例,其中 A=被错误诊断的患者数,B=患者总数。									
经济损失	经济损失的影响范围怎样?	使用统计	$X=1-A/B$ A=发生经济损失的次数 B=使用总数	$0.0 \leqslant X \leqslant 1.0$ 越趋近于 1.0 越好	绝对标度	A=计数 B=计数 X = 计数/计数	使用监控记录	5.4 运作	用户 人机界面设计者 开发者
注:这种度量也可由具有经济损失风险的情况的发生次数来测量。									
软件损坏	软件讹误(corruption)的影响范围怎样?	使用统计	$X=1-A/B$ A=发生软件讹误的次数 B=使用总数	$0.0 \leqslant X \leqslant 1.0$ 越趋近于 1.0 越好	绝对标度	A=计数 B=计数 X = 计数/计数	使用监控记录	5.4 运作	用户 人机界面设计者 开发者
注 1:这种度量也可由具有软件损坏风险的情况的发生次数来测量。 注 2:也可用另一种测量方法:X=软件讹误引起的累积成本/使用时间。									

表 8.4 满意度度量

度量名称	度量目的	应用的方法	测量、公式及数据元素计算	测量值解释	度量标度类型	测度类型	测量输入	在 GB/T 8566 中的应用	目标用户
满意度标度	用户满意程度?	用户测试	$X=A/B$ A=通过调查问卷得到的心理测试标度 B=人口总体平均数	$0.0<X$ 越大越好	比率标度	A=计数 X=计数	运行(测试)报告 用户监视记录	6.5 确认 5.3.9 合格性测试 5.4 运作	用户 人机界面设计者
注：心理测试调查问卷的范例参见 F.3。									
满意度问卷	用户对具体的软件特征的满意程度怎样?	用户测试	$X=\Sigma(A_i)/n$ A_i=对问题的响应 n=响应数	与前面得到的值相比较或与总体平均数比较	顺序标度	A=计数 X=计数	运行(测试)报告 用户监视记录	6.5 确认 5.3.9 合格性测试 5.4 运作	用户 人机界面设计者 开发者
注：如果由问卷中的各项综合给出一个总的记分，由于不同的问题可能有不同的重要性，应该对其进行加权。									
选用度	选择使用该系统的潜在用户的比例是多少?	使用观察	$X=A/B$ A=使用特定软件功能/软件应用/软件系统的次数 B=打算使用它们的次数	$0.0\leqslant X\leqslant 1.0$ 越趋近于 1.0 越好	比率标度	A=计数 B=计数 X=计数/计数	运行(测试)报告 使用用户监视记录	6.5 确认 5.3.9 合格性测试 5.4 运作	用户 人机界面设计者 开发者
注：这种度量适用于任意使用的情况。									

附 录 A
（资料性附录）
使用度量时的考虑

A.1 测量的解释

A.1.1 测试使用周境与运行使用周境之间的潜在差异

在策划使用度量或解释测度时，理解清楚软件所要的使用周境，以及测试使用周境与运行使用周境之间的潜在差异是很重要的。例如：在类似的软件系统中，“学会操作所需的时间”的测度，对技术熟练的操作者与非熟练的操作者常常不一样。下面给出一些潜在差异的例子：

a） 测试环境与运行环境之间的差异

测试环境与运行环境之间是否有明显差异？

下面是一些实例：

- 具有较高/相当/较低的运行计算机 CPU 性能的测试环境；
- 具有较高/相当/较低的运行网络和通信性能的测试环境；
- 具有较高/相当/较低的运行操作系统性能的测试环境；
- 具有较高/相当/较低的运行用户界面性能的测试环境。

b） 测试的执行与实际运作的执行之间的差异

测试的执行与用户环境中运行的执行之间是否有明显差异？

下面是一些实例：

- 测试环境中功能的覆盖率；
- 测试用例的抽样率；
- 实时事务的自动测试；
- 压力负载；
- 每周 7×24 h(不间断)运行；
- 用来测试异常和差错的合适数据；
- 周期性处理；
- 资源利用率；
- 中断级别；
- 生产强度；
- 干扰。

c） 观察用户的特征

测试时用户的特征与运行时用户的特征是否有明显差异？

如下列实例：

- 混合类型的用户；
- 用户的技能水平；
- 专业用户或一般用户；
- 受限用户组或公共用户。

A.1.2 影响结果有效性的问题

下列问题可能会影响所收集的数据的有效性。

a） 收集评价结果的规程

- 借助工具或设施自动收集/手工收集/问卷调查或面谈。

b) 评价结果的来源

- 开发者的自述报告/评审者的报告/评价者的报告。

c) 结果数据的确认

- 开发者自查/由独立评价者检查。

A.1.3 测量资源的平衡

在每个阶段使用的测度的平衡是否适合于评价的目的?

在为内部测量、外部测量和使用质量的测度应用适当的度量范围时,平衡所用的工作量是很重要的。

A.1.4 规格说明的正确性

软件规格说明与实际操作要求之间是否有明显差异?

在不同阶段评价软件产品时,所采取的测量就是与产品的规格说明进行比对。因此,通过确认和验证来确保用于评价的产品规格说明能反映出运行中真实和实际的需要是非常重要的。

A.2 度量的确认

A.2.1 度量的理想性质

为了从质量评价中获得有效结果,度量应具有下列性质。若某种度量不具备这些性质,那么,度量描述应说明对其有效性的约束,并尽可能解释如何处理这类情况。

a) (度量的)可靠性:可靠性与随机误差有关。如果随机变量不影响度量的结果,则度量是没有随机差错的。

b) (度量的)可重复性:由相同的评价者使用相同的评价规格说明(包括在相同的环境中)和相同的用户类型及环境,对相同产品重复进行的度量宜在适当的容差范围内得出相同的结果。这里所谓适当的容差应包括诸如疲劳、学习效应等因素。

c) (度量的)可再现性:由不同的评价者使用相同的评价规格说明(包括在相同的环境中)和相同的用户类型及环境,对相同产品进行的度量宜在适当的容差范围内得出相同的结果。

注 1:建议对测量结果的可变性进行统计分析。

d) (度量的)可用性:度量应明确指出其约束使用条件(如特定因素存在的条件)。

e) (度量的)指示性:度量标识软件应改进的部份或改进的项,并给出与期望值进行比较的测量结果的能力。

注 2:与只检查所需的项目不同,对选定或建议的度量宜提供使用度量可用性的书面证据。

f) (测度的)正确性:度量宜应具备下列性质:

1) (测度的)客观性:度量的结果与其数据输入应是有据可查的,即不受评价者、测试用户的感觉或观点的影响(除非满意度或吸引性度量,因为用户的感觉与观点也是测量的对象)。

2) (测度的)公正性:度量不应偏向任何特殊的结果。

3) (测度的)充分精确性:精确性由度量的设计,特别是作为度量基础的材料的选择来确定。度量的用户将描述度量的精确性和灵敏性。

g) (测度的)意义:测量应产生有关软件行为或质量特性的有意义的结果。度量也应具有成本效益:即成本越高的度量,提供的结果应越具价值。

A.2.2 度量有效性的证实

度量的用户应标识一些证实度量的有效性的方法,例如:

a) 相关性

质量特性值(对运行使用中的主要度量的测度)中的变化可以用度量值中的变化来解释,用线性系数的平方表示。

利用相关性度量,评价者不用直接测量就可以预测质量特性的值。

b) 跟踪

若度量值 M 和质量特性值 Q(对运行使用中的主要度量的测度)直接相关,给定一个产品或过程,当值 $Q(T_1)$变为 $Q(T_2)$时,度量值也以相同的趋势,从 $M(T_1)$变为 $M(T_2)$(若 Q 值增加,则 M 值也增加)。

评价者不必直接测量而是通过使用那些具有跟踪能力的度量就可以检测质量特性随时间周期的变化。

c) 一致性

若质量特性值(对运行使用中的主要度量的测度)$Q_1, Q_2, \ldots, Q_n$ 对应于产品或过程 $1, 2, \cdots, n$,当有关系 $Q_1 > Q_2 > \cdots > Q_n$ 时,则对应的度量值也有关系 $M_1 > M_2 > \cdots > M_n$。

评价者可以使用具有一致性能力的度量来关注软件的异常部件和易出差错的部件。

d) 可预测性

若使用时间 T_1 的度量来预测时间 T_2 质量特性值 Q(对运行使用中的主要度量的测度),则预测误差值应在预测允许的范围内。预测误差=(预测值 $Q(T_2)$-实际值 $Q(T_2)$)/ 实际值 $Q(T_2)$。

评价者可以通过可预测性的度量预测质量特性今后的变化趋势。

e) 可判别性

度量能够判别出软件质量的高低。

评价者可以使用具有判别能力的度量对软件部件进行分类和对质量特性值定级。

A.3 使用度量进行估计(判断)与预测(展望)

用如下两种方法在早期阶段估计和预测软件产品的质量特性是最具价值的度量。

A.3.1 利用当前的数据预测质量特性

a) 利用回归分析来预测

当通过使用特性(属性)的当前值(数据)来预测同一特性(属性)的未来值(测度)时,根据一个足够长的时间内观察到的一组数据进行回归分析是有用的。

例如在测试阶段(活动)中获得的 MTBF(平均失效间隔时间)的值可用于估计在运行阶段的 MTBF。

b) 利用相关性分析来预测

当用不同属性的当前测量值来预测特性(属性)的未来值(测度)时,使用一个确认的表明相关性的函数进行相关性分析是有用的。

例如在编码阶段,模块的复杂性可用来预测在维护过程中程序修改和测试所花费的时间与工作量。

A.3.2 根据当前的事实估计当前的质量特性

a) 利用相关性分析来估计

在估计不可直接测量的属性的当前值时,若任何其他的测度与目标测度密切相关时,相关性分析方法是有用的。

例如软件产品中遗留的故障数是不可测的,但它可以用检测出的故障数及故障趋势进行估计。

对于不能直接测量的属性进行预测的那些度量应用下述解释来估计:

- 使用模型来预测属性;
- 使用公式来预测属性;
- 基于经验来预测属性;
- 使用合理判断来预测属性。

对于不能直接测量的属性进行预测的那些度量可以用下述解释来确认:

- 标识要预测的属性的测度;
- 标识要用来预测的度量;

- 进行基于确认的统计分析；
- 将结果归档；
- 定期地重复上述工作。

A.4 检测易发生质量问题的部件中的偏差或异常

下列质量控制工具可用来分析在软件产品部件中的偏差和异常情况：

a) 流程图(软件的功能模块)；

b) 排列分析和排列图；

c) 直方图和散点图；

d) 运行图、相关图和层次图；

e) 鱼骨图；

f) 统计过程控制(软件功能模块)；

g) 检查单。

上述工具可用于标识源于数据的质量问题，这些数据是通过应用度量来获得的。

A.5 显示测量结果

a) 显示质量特性评价的结果

对于每个质量特性和子特性可用下列图示法显示质量评价的结果：

雷达图、条形图、数字化的直方图、多变量图、重要性能矩阵图等。

b) 显示测度

可利用一些有用的图形表示，如排列图、趋势图、直方图、相关图等。

附 录 B
（资料性附录）
使用质量的度量、外部度量和内部度量的用法（框架实例）

B.1 引言

本框架实例是一个高层描述，它描述的是关于如何在软件开发和实现过程中使用GB/T 16260.1—2006中的质量模型和相关的度量来获得满足用户要求的质量产品。本实例所示的概念可用不同的定制形式来实现，以适用于个体、组织或者项目。本实例使用的GB/T 8566—2001中的生存周期过程可作为传统软件开发生存周期的基准，使用的GB/T 18905.3—2002中的质量评价过程的步骤可作为传统软件产品质量评价过程的基准。只要能理解基本概念，如果用户愿意，也可以把这些概念映射为其他的软件生存周期模型。

B.2 开发及质量过程的概述

为了测量可交付项（即使用质量、外部质量和内部质量）的质量，表B.1描述了一个示例模型，它把软件开发生存周期过程的活动（从活动1到活动8）与其关键的可交付项及相关的引用模型联系起来。

第一行描述软件开发生存周期过程的活动。（可为适应独特的要求来定制）。第二行描述可能作为测量类别（如使用质量，外部质量或内部质量）的一种实际的测度或者预测。第三行描述可以测量质量的关键可交付项，第四行描述在每个过程活动中可适用于每个可交付项的度量。

表 B.1 质量测量模型

	活动1	活动2	活动3	活动4	活动5	活动6	活动7	活动8
阶段	需求分析（软件与系统）	体系结构设计（软件与系统）	软件的详细设计	软件编码与测试	软件集成及软件的合格性测试	系统集成及系统合格性测试	软件的安装	软件的验收支持
模型的引用	所需的用户质量 所需的内部质量 所需的外部质量	预测的使用质量， 预测的外部质量， 测量的内部质量	预测的使用质量， 预测的外部质量， 测量的内部质量	预测的使用质量， 测量的外部质量， 预测的外部质量， 测量的内部质量	预测的使用质量， 测量的外部质量， 预测的外部质量， 测量的内部质量	预测的使用质量， 测量的外部质量， 测量的内部质量	预测的使用质量， 测量的外部质量， 测量的内部质量	测量的使用质量， 测量的外部质量， 测量的内部质量
活动的关键可交付项	用户的质量需求（规定的）， 外部的质量需求（规定的）， 内部的质量需求（规定的）	软件/系统体系结构设计	软件详细设计	软件代码，测试结果	软件产品，测试结果	集成的系统，测试结果	安装的系统	交付的软件产品
用于测量的度量	内部度量（外部度量可用于确认需求规格说明）	内部度量	内部度量	内部度量 外部度量	内部度量 外部度量	内部度量 外部度量	内部度量 外部度量	使用质量度量 内部度量 外部度量

B.3 质量途径步骤

B.3.1 概述

开发周期中对质量的评价分为下列步骤。步骤1必须在需求分析活动中完成，步骤2到5必须在上述定义的每个过程活动中重复进行。

B.3.2 步骤1：质量需求的确定

对质量模型中定义的每个质量特性和子特性，用表B.2中的两个例子对每类测量(使用质量、外部和内部质量)确定用户要求的权重。根据分配的相对权重，允许评价者将精力集中在最重要的子特性上。

表 B.2 用户要求的特性与权重a)

使用质量		
	特性	权重(高/中/低)
	有效性	高
	生产率	高
	安全性	低
	满意度	中

表 B.2 用户要求的特性与权重 b)

外部与内部质量		
特性	子特性	权重(高/中/低)
功能性	适合性	高
	准确性	高
	互操作性	低
	安全保密性	高
	功能性的依从性	中
可靠性	成熟性(硬件/软件/数据)	低
	容错性	低
	易恢复性(数据、过程、技术)	高
	可靠性的依从性	高
易用性	易理解性	中
	易学性	低
	易操作性	高
	吸引性	中
	易用性的依从性	高
效率	时间特性	高
	资源利用性	高
	效率的依从性	高

表 B.2 b)(续)

外部与内部质量		
特性	子特性	权重(高/中/低)
维护性	易分析性	高
	易改变性	中
	稳定性	低
	易测试性	中
	维护性的依从性	高
可移植性	适应性	高
	易安装性	低
	共存性	高
	易替换性	中
	可移植性的依从性	高

注：权重可用高/中/低的方式表示，也可在 1～9 的范围内用顺序标度来表示(例如 1～3＝低、4～6＝中、7～9＝高)。

B.3.3 步骤 2：评价的规格说明

每个开发过程活动都要实施本步骤。

质量模型中定义的每个质量子特性均标识要应用的度量和要求的级别，以便达到在第 1 步中设定的用户要求，并按表 B.3 的例子加以记录。

对内容阐述的基本输入及用法说明可在表 B.1 的例子中得到，其中解释了在开发周期的这一阶段中能测量什么。

注：在开发周期的特定活动中，表中的某些行可以是空的，因为在开发过程的早期，不可能测量所有子特性。

表 B.3 质量测量表 a)

使用质量测量类别				
	特性	度量	要求的级别	评估实际结果
	效率			
	生产率			
	安全性			
	满意度			

表 B.3 质量测量表 b)

外部质量测量类别				
特性	子特性	度量	要求的级别	评估实际结果
功能性	适合性			
	准确性			
	互操作性			
	安全保密性			
	功能性的依从性			

表 B.3 b)(续)

外部质量测量类别				
特性	子特性	度量	要求的级别	评估实际结果
可靠性	成熟性(硬件/软件/数据)			
	容错性			
	易恢复性(数据、过程、技术)			
	可靠性的依从性			
易用性	易理解性			
	易学性			
	易操作性			
	吸引性			
	易用性的依从性			
效率	时间特性			
	资源利用性			
	效率的依从性			
维护性	易分析性			
	易改变性			
	稳定性			
	易测试性			
	维护性的依从性			
可移植性	适应性			
	易安装性			
	共存性			
	易替换性			
	可移植性的依从性			

表 B.3 质量测量表 c)

内部质量测量类别				
特性	子特性	度量	要求的级别	评估实际结果
功能性	适合性			
	准确性			
	互操作性			
	安全保密性			
	功能性的依从性			
可靠性	成熟性(硬件/软件/数据)			
	容错性			
	易恢复性(数据、过程、技术)			
	可靠性的依从性			

表 B.3 c)(续)

内部质量测量类别				
特性	子特性	度量	要求的级别	评估实际结果
易用性	易理解性			
	易学性			
	易操作性			
	吸引性			
	易用性的依从性			
效率	时间特性			
	资源利用性			
	效率的依从性			
维护性	易分析性			
	易改变性			
	稳定性			
	易测试性			
	维护性的依从性			
可移植性	适应性			
	易安装性			
	共存性			
	易替换性			
	可移植性的依从性			

B.3.4 步骤 3:评价的设计

每个开发过程活动都要实施本步骤。

制定一个包括可交付项的测量计划(类似于表 B.4 中的例子),这些交付项用作要实施的度量和测量过程的输入。

表 B.4 测量计划

子特性	要评价的可交付项	应用的内部度量	应用的外部度量	应用的使用质量度量
1. 适合性	1. 2. 3.	1. 2. 3.	1. 2. 3.	(不适用)
2. 满意度	1. 2. 3.	(不适用)	(不适用)	1. 2. 3.
3.				
4.				
5.				
6.				

B.3.5 步骤 4:评价的执行

每个开发过程活动都要实施本步骤。

执行评价计划，填写表 B.3 实例中的每一列。GB/T 18905 系列标准可用作编制计划和执行测量过程的指南。

B.3.6 步骤 5：反馈给组织

每个开发过程活动都要实施本步骤。

一旦所有测量均已完成，要把结果映射到表 B.1 中并以报告的形式将结论写成文件。同时标识产品质量需要改进的特定区域，以使其满足用户的需要。

附 录 C
（资料性附录）
度量标度类型和测量类型的详细解释

C.1 度量标度的类型

当度量的用户得到了测量结果并使用该测度进行计算和比较时，宜对每种测度标识下列度量标度类型。对某些测度，平均值、比率或差值可能没有意义。度量标度类型有：标称标度、顺序标度、间隔标度、比率标度和绝对标度。$M'=F(M)$，这里 F 是一个容许函数。每个测量标度类型的描述包含容许函数的描述（若 M 是一个度量，则 $M'=F(M)$ 也是一个度量）。

a） 标称标度

$M'=F(M)$，这里 F 是一对一的映射。

标称标度包括分类，例如软件的故障类型（数据、控制、其他）。只有用相同类型的频率计算时，平均值才有意义。只有用经过映射的每种类型的频率计算时，比率也才有意义。因此，平均值和比率可以用来代表相同类型频率的早期和后来情况之间或两个类似情况之间的差。否则，它们可以用来相互比较每种类型各自的频率。

注 1：例如：城市交通线标识号、编译器出错消息标识号。

注 2：含义说明：只是不同类别的数。

b） 顺序标度

$M'=F(M)$，这里 F 是一个单调递增的映射，即：若 $M(x)\geqslant M(y)$，则 $M'(x)\geqslant M'(y)$。

顺序标度包括排序，例如：软件的失效按严重程度排序（忽略不计的、轻微的、严重的、灾难性的）。只有用经过映射的相同顺序的频率计算时，平均值才有意义。只有用经过映射的每种顺序的频率计算时，比率也才有意义。因此，比率和平均值可用来代表相同顺序频率的早期和后来情况之间或两个类似情况之间的差。否则，它们可以用来相互比较每种顺序的频率。

注 1：例如：学校的考试成绩：优、良、及格和不及格。

注 2：含义说明：每个量取决于它们在顺序中的位置，如中值。

c） 间隔标度

$M'=aM+b\ (a>0)$

当两次测度之间的差值有经验意义时，间隔标度包括排序的等级标度。但间隔标度中两次测度的比率可能没有相同的经验意义。

注 1：例如：温度（摄氏、华氏、开氏），实际计算时间与预测的时间的差。

注 2：含义说明：算术平均值和任何依赖排序的值。

d） 比率标度

$M'=aM\ (a>0)$

当两次测度之间的差值及两次测度的比例有相同的经验意义时，比率标度包括排序的等级标度。平均值和比率有各自的含义，它们给出了值的实际含义。

注 1：例如：长度、重量、时间、规模、计数。

注 2：含义说明：几何平均、百分比。

e） 绝对标度

$M'=M$，它们只能按一种方式测量。

任何与测度有关的说明都是有意义的。例如，当测量的单位相同时，一个比率标度类型的测度除以另一个比率标度类型的测度，结果是一个绝对值。一个绝对标度类型的测量值事实上不带任何单位。

注 1：例如：带注释的代码行数除以代码的总行数。

注 2：含义描述：一切事情。

C.2 测度类型

C.2.0 概述

为了设计一个收集数据、正确解释其含义并且把测度规范化以便进行比较的过程，度量的用户宜标识并考虑度量所使用的测度类型。

C.2.1 规模测度类型

C.2.1.0 导引

按其定义中所声称的测度内容，本类型的测度代表软件的一种特殊规模。

注：软件可以有多种表示规模的方法(就像任何一个实体可以进行多维测量——质量、体积、表面积等等。)

用一种规模测度来使其他的测度规范化，可以根据规模单位给出可比值。下列描述的规模测度类型可用于软件质量的测量。

C.2.1.1 功能规模类型

功能规模是软件可能有的一种规模类型(一维)的例子。任何一个软件实例可能会有多个功能规模，例如取决于：

a) 测量软件规模的目的(它影响到在测量中包含的软件范围)；

b) 所用的特定功能规模测量方法(它将改变其单位和标度)。

GB/T 18491.1—2001 提供了概念定义和使用功能规模的测量方法(FSM 方法)的过程。

为了规范化地使用功能规模，必须确保采用相同的功能规模方法，基于同样的目的，还要确保要比较的不同软件已经过测量，因而具有可比较的范围。

尽管下列内容经常声称代表了功能规模，但不能保证它们等同于应用 FSM 方法所获得的功能规模，也不能保证它们依从于 GB/T 18491.1—2001。不过，在软件开发中，如下的方法仍被广泛使用。

1) 电子表格数；

2) 屏幕数；

3) 要处理的文件或数据集合数；

4) 用户需求规格说明描述的逐条列举的功能需求数。

C.2.1.2 程序规模类型

本条中，术语“程序设计”代表当执行时导致一些动作的表达式，术语“语言”代表所用的表达式类型。

1. 源程序规模

应解释程序设计语言，它应提供如何处理诸如注释行这样的不可执行语句。经常使用下列测度：

非注释性源语句(NCSS)包括可执行语句和带有逻辑性源语句的数据声明语句。

注 1：新程序规模：

开发者可能使用新开发的程序规模来代表开发与维护工作产品的规模；

注 2：变更的程序规模：

开发者可能使用变更的程序规模来代表包含修改过的部件的软件规模；

注 3：计算的程序规模：

计算程序规模的公式的例子：新代码行 + 0.2 * 修改过的部件中的代码行(NASA Goddard)

可能有必要更详细地区分下列源代码语句的类型：

Ⅰ. 语句的类型

逻辑性源语句(LSS)：LSS 测量软件指令的数量。这些语句不考虑与行的关系，独立于表现它们的物理格式。

物理源语句(PSS)：PSS 测量软件的源代码行数。

Ⅱ. 语句的属性

可执行语句；

数据声明语句；

编译程序命令语句；

注释性源语句。

Ⅲ. 源

修改的源语句；

增加的源语句；

删除的源语句。

- 新开发的源语句(= 增加的源语句+修改的源语句)；
- 重用的源语句(= 原来的源语句－修改的源语句 － 删除的源语句)。

2. 程序字计数规模

可采用下列 Halstead 测度方法计算测量值：

程序的词汇数 $= n_1 + n_2$；观察到的程序长度 $= N_1 + N_2$。其中：

- n_1：程序源代码中被程序语言预留的不同操作符的字数；
- n_2：程序源代码中由编程人员定义的不同操作数的字数；
- N_1：程序源代码中不同操作符出现的次数；
- N_2：程序源代码中不同操作数出现的次数。

3. 模块数

本测量计算可独立执行的对象个数，例如程序中的模块个数。

C.2.1.3 利用的资源规模测度类型

本测度类型标识要评价的软件在运行中所用的资源。例如：

a) 存储器的数量，例如在软件执行过程中，临时和永久占用的磁盘或内存的数量；

b) I/O 负载，例如通信数据流量总数(对网络中的备份工具有意义)；

c) CPU 负载，例如每秒钟 CPU 指令集占用的百分比(本测度类型对测量 CPU 的利用率或在并发/并行系统中软件的多线程运行时测量进程分配的效率时有意义)；

d) 文件与数据记录，例如文件或记录的位长度；

e) 文档，例如文档的页数。

注意峰值(最大值)、最小值和平均值，以及时间周期及观察的次数等数据可能很重要。

C.2.1.4 特定的操作规程步骤类型

本测度类型标识在人工界面的设计规格说明或用户手册中规定的规程的静态步骤。本测量值可能依测量所用的描述类型的不同而有所区别，例如用户的操作规程可以用图形也可以用文字来表示。

C.2.2 时间测度类型

C.2.2.0 导引

时间测度类型度量的用户应记录时间周期、检查过多少站点及有多少用户参与了这一测量。有多种以时间为单位进行测量的方式，例如：

a) 实时单位

这是物理时间单位，如秒、分或小时。这种单位常用来描述实时软件的任务处理时间。

b) 计算机机器时间单位

这是计算机处理器的时钟时间，即 CPU 时间的秒、分或小时。

c) 正式的日程表上的时间单位

包括工作小时、日历(日、月或年)。

d) 部件的时间单位

在有多个站点时，部件时间单位标识各个站点，部件时间单位是每个站点单独时间的累计。这种单位通常用来描述部件的可靠性，如部件的失效率。

e) 系统时间单位

在有多个站点时，系统时间不标识单独的站点，而标识整个系统中所有运行的站点。这种单位常用来描述系统的可靠性，如系统的失效率。

C.2.2.1 系统运行时间类型

系统运行时间类型为测量软件的可用性提供了基础。主要用于评价可靠性。应确定软件是间断运行还是连续地运行。如果软件是间断运行的，应确保在软件运行期间对时间进行测量(这显然可以扩展到连续运行的情况)。

a) 经时时间

当软件在不变的情况下使用时，如系统每周运行时间长度相同。

b) 机器加电时间

用于实时的、嵌入的或操作系统软件，它在系统运行的全部时间内都得到充分使用。

c) 规格化的机器时间

类似于机器加电时间，但把多台机器上不同的加电时间数据汇集起来并用一个修正因子进行调整。

C.2.2.2 执行时间类型

执行时间类型是指为完成特定任务所需要执行软件的时间。应分析几种尝试的分布，应计算均值、方差和最大值。应检查在特定条件下，特别是在过载条件下的执行时间。执行时间类型主要用于评价效率。

C.2.2.3 用户时间类型

用户时间类型测量单个用户在使用软件完成任务时所花费的时间。例如：

a) 会话时间

会话开始和结束的时间。如一个家庭银行系统的用户提取钱的行为。对于交互程序来说，只研究交互的易用性问题，不研究空闲时期。

b) 任务时间

单个用户每次试图运行软件完成任务所花费的时间。应定义好测量的起点和终点。

c) 用户时间

从开始到某个时间点，用户使用软件所花费的时间(从开始时起，用户使用软件大约有多少小时的时间或天数)。

C.2.2.4 工作量类型

工作量类型是指与某特定项目任务有关的生产时间。

a) 个人工作量

开发者、维护者或操作者为完成特定任务进行工作所需要的生产时间。个人的工作量只是每天一定数量的生产小时数。

b) 任务工作量

任务工作量是指所有单个的项目人员(开发者、维护者、操作者、用户或其他)为完成特定任务进行工作的人员工作量的累计值。

C.2.2.5 事件的时间间隔类型

本测度类型是指在观察期间，一个事件与下一个事件之间的时间间隔。可用观察时段的频率代替本测度。本测度可以典型地用来描述相继发生的失效之间的时间。

C.2.3 计数测度类型

C.2.3.0 导引

若对软件产品的文档属性进行计数，则为静态计数类型。若对事件或人的动作进行计数，则为动态

计数类型。

C.2.3.1 检测的故障数类型

本测量对在评审、测试、纠正、运行或维护期间检测到的故障个数进行计数。按照故障所造成的影响,可为这些故障的严重程度进行分类。

C.2.3.2 程序结构的复杂度类型

本测量对程序结构的复杂度进行计数。例如不同路径的数目或 McCabe 圈复杂度。

C.2.3.3 检测不一致的个数类型

本测度对调查所准备的不一致项数进行计数。

a) 不符合的项数

例如:

- 与需求规格说明的规定项不相符;
- 与法律、法规或标准不相符;
- 与协议、数据格式、介质格式、字符编码不相符。

b) 用户期望的不能实现实例数

本测量对所列举的满意或不满意的项数进行计数,这些项描述用户合理的期望与软件产品性能间的差别。

本测量可用问卷的方式向测试者、客户、操作者或最终用户就发现的缺陷进行调查。例如:

- 功能是否可用;
- 功能是否有效地执行;
- 功能是否可用于用户特定的预期使用;
- 功能是否是预期的、需要的或不需要的。

C.2.3.4 变更数类型

本类型标识检测出的已经变更的软件配置项。如在源代码中发生变更的行数。

C.2.3.5 检测到失效数类型

本测量对在产品开发、测试、运作或维护过程中检测出的失效个数进行计数。根据这些失效造成的影响,可以按严重性的等级进行分类。

C.2.3.6 尝试(试验)次数类型

本测度对与故障造成的缺陷相关的尝试次数进行计数。例如在评审、测试和维护中的尝试次数。

C.2.3.7 人工操作过程中的点击类型

当用户与软件在操作中发生互动时,本测度对用户行为的动态步骤活动所产生的点击个数进行计数。本测度量化了人类工效的易用性及使用的工作量。因此,本测度可用于易用性测量。如执行任务时的点击次数,眼睛活动的次数等。

C.2.3.8 记分类型

本类型标识算术计算的记分或结果。记分可包括计数或按检查表进行或不进行加权计算。例如检查表的记分;问卷调查的记分;Delphi 方法等。

附 录 D
（资料性附录）
术 语

D.1 定义

除非特别指出，全部的定义都引自 GB/T 18905.1—2002 和 GB/T 16260.1—2006。

D.1.1 质量 quality

外部质量 external quality

产品在特定条件下使用时，满足明确或隐含要求的程度。

内部质量 internal quality

产品属性的总和，决定了产品在特定条件下使用时，满足明确和隐含要求的能力。

注 1：当术语“特性”在本部分中用于更特定的意思时，使用术语“属性”（而不是 3.1.3 中使用的术语“特性”）。

质量 quality

实体特性的总和，表示实体满足明确或隐含要求的能力。

注 2：在某种契约的环境或在某个受控的环境中，如核安全领域，要求是明确规定的，而在其他环境中，宜确定和定义隐含的要求。

使用质量 quality in use

特定用户使用产品满足其要求的程度，以达到在特定应用环境中的有效性、生产率和满意度等特定目标。

注 3：使用质量是在包含软件的环境中质量的用户观点，它可以用在环境中使用软件的结果来测量，而不是根据软件本身的性质来测量。

注 4：在 GB/T 18905.1—2002 中使用质量的定义目前不包括新的“安全性”特性。

质量模型 quality model

一组特性及特性之间的关系，它提供规定质量需求和评价质量的基础。

D.1.2 软件与用户 software and user

软件 software

信息处理系统的部分或全部程序、过程、规则及相关的文档。[GB/T 5271.1—1993]。

注 1：软件是独立于所记录媒体的智力创作。

软件产品 software product

一组计算机程序、规程以及可能有的相关文档和数据。[GB/T 8566—2001]

注 2：产品包括中间产品及专为开发者或维护者这样的用户所准备的产品。

用户 user

使用软件产品执行特定功能的个人。

注 3：用户可以包括操作者、软件结果的接受者或软件的开发者或维护者。

D.1.3 测量 measurement

属性 attribute

实体的可以测量的物理或理论上的性质。

直接测度 direct measure

不依赖于任何其他属性测度的一种对属性的测度。

外部测度 external measure

从对系统行为的测度导出的对产品的一种间接测度，其中产品是系统的一部分。

注 1：系统包括任何相关的硬件、软件(定制的软件或现货软件)和用户。

注 2：在测试中发现的故障数量是对程序中的故障数量的外部测度，因为故障的数量是在计算机系统运行程序的过程中计数，以便标识代码中故障的数量。

注 3：外部测度可以用来评价更接近于最终设计目标的质量属性。

指标 indicator

能用来估计或预测另一测度的一种测度。

注 4：预测的测度可以是针对相同或不同的软件质量特性。

注 5：指标可用来估计软件质量的属性和开发过程的属性，它们是对属性的间接测度。

间接测度 indirect measure

从一个或一个以上的其他属性的测度导出的一种对属性的测度。

注 6：对计算机系统属性(例如对用户输入的响应时间)的外部测度就是对软件属性的一种间接测度，因这种测度要受计算环境的属性和软件属性的影响。

内部测度 internal measure

对产品本身的一种测度，或是直接的或是间接的。

注 7：代码行数、复杂度、在走查和 Fog 索引中发现的故障数都是对产品本身进行的内部测度。

测度(名词) measure(noun)

通过进行一次测量赋予实体属性的数或类别。

测量(动词) measure(verb)

进行一次测量。

测量(名词) measurement(noun)

使用一种度量，把标度值(可以是数或类别)赋予实体的某个属性。

注 8："类别"可用于指示属性的定性测度。如软件产品的一些重要属性，例如源程序语言(Ada，C，COBOL 等)就是定性的类别。

度量 metric

定义的测量方法和测量标度。

注 9：度量可以是内部的或外部的。

度量包括对定性数据进行分类的方法。

附 录 E
（资料性附录）
使用质量的评价过程

E.1 确定评价需求

本章遵循 GB/T 18905.1—2002 中的评价过程结构。

E.1.1 确定评价目的

评价使用质量的目的是评估在特定使用周境（使用场景）中，为达到特定目标，产品满足用户要求的程度。

E.1.1.1 获取

开发前，为了确定产品是否满足使用质量需求，一个组织寻求获取适合其需要的产品时，可以将使用质量作为框架，用以确定产品应该满足的以及经验收测试而不能满足的使用质量需求。应确定测量使用质量的特定使用周境，选择有效性、生产率、安全性及满意度的测度并确立基于这些测度的验收准则。

E.1.1.2 供给

供方通过评价使用质量以确保产品满足特定用户类型和特定使用环境的需要。给潜在的需方提供使用质量的结果将有助于需方判断产品是否符合特定需要。（见附录 F，附录 G 中的例子）

E.1.1.3 开发

对不同使用场景下的使用质量来说，对用户需求的清晰理解将有助于开发团队针对满足真实的用户要求去进行设计决策，把开发目标集中在满足使用质量的准则上。在软件开发完成时，以评价这些准则。

E.1.1.4 运作

运行系统的组织可以通过测量使用质量来评价系统满足其要求的程度，及评估在未来的版本中可能需要的变更。

E.1.1.5 维护

对于软件维护者，可以测量维护任务的使用质量。对于软件移植者，可以测量移植任务的使用质量。

E.1.2 确定产品类型

为了评价使用质量，需要一个工作原型或最终产品。

E.1.3 规定质量模型

使用质量所选用的质量模型是 GB/T 16260.1—2006 中的质量模型。其中，使用质量是指：在特定的使用周境中，产品使特定用户达到特定目标所要求的有效性、生产率、安全性及满意度的能力。

E.2 规定评价

E.2.1 标识使用周境

为了规定或测量使用质量，必须标识使用周境各个组成部分，如用户、目标及使用的环境。通常，要测试所有可能的使用周境是不可能的。因此，有必要选择重要的或有代表性的用户组和任务。

E.2.1.1 用户

需要规定用户在使用产品时影响用户能力的特征。包括知识、技能、经验、教育、培训、体格特性、驱动力和感官能力。必须定义不同类型用户的特征，例如，具有不同经验水平和扮演不同角色的用户

E.2.1.2 目标

应规定产品的使用目标。规定要达到的目标，而不是如何达到目标。目标可以分解成各子目标及

满足各子目标的准则，这些子目标构成了总目标。例如，如果目标是要完成客户订单，子目标可能就是在每个字段输入正确信息。总目标的大小依赖于评价范围。任务是为了完成目标所需要的活动。

E.2.1.3 环境

运行环境

应规定硬件和软件运行环境，因为运行环境可能会影响软件执行的方式。包括网络响应时间等更广泛的内容。

用户环境

应规定影响用户能力的工作环境的各个方面，例如物理环境（如工作场所、设备）、野外环境（如温度、光照）及社会文化环境（工作实践，获取帮助的渠道及动机）。

E.2.2 选择评价的周境

重要的是评价使用的周境要尽可能和产品实际使用环境相接近。用来预测产品在实际使用时使用质量所达到级别的测度的有效性依赖于用户、任务、环境和实际环境相接近的程度。一个极端的情况是在野外可使用实际的工作情况作为评价产品使用质量的基础。另一方面，也可以在以一种典型的和可控制的方式重建的与使用周境相关的实验室环境中来评价产品的特定方面。实验室中评价的优势是这种方法可以对影响使用质量水平的关键变量实施更多的控制，并可以进行更精确的测量。缺点就是实验环境是人为制造的，可能会产生不现实的结果。

E.2.3 选择度量

E.2.3.1 选择测度

为了规定或评价使用质量，通常必须对有效性、生产率、满意度及相关的安全性至少测量一种度量。

度量及其测量周境的选择依赖于测量所涉及的各方的目标。应考虑每种度量对目标的相对重要性。如对不常使用的度量，则应更注重它的易理解性和易学性而不是使用质量方面。

使用质量的测度应基于数据。这些数据反映了用户与产品交互的结果。利用客观手段采集数据是可能的。如输出，工作速度及特定事件发生率的测量。另外，也可以对用户表达的情感、信仰、态度或偏好等主观反应采集数据。客观测度直接表明了有效性和生产率，而主观测度则直接与满意度相关。

根据需要研究的问题和可用以测试产品的完备性，可以在现场和实验室等不同地点进行评价。测度和测试环境的选择依赖于测量活动的目标及其与设计周期的关系。

E.2.3.2 有效性

有效性度量是用来测量目标可以达到的准确性和完备性。

例如，如果期望的目标是依据规定的格式准确地重新生成两页文档，那么准确性就可以通过拼写错误数、与规定格式相偏离数来规定或测量，完备性可用转录文档中的字数除以源文档中的字数来规定或测量。

E.2.3.3 生产率

生产率测度将达到的有效性级别与所消耗的资源相关联。相关的资源包括智力、体力、时间、材料和财力。如，有效性/人力＝人的生产率。有效性/时间＝时间生产率，有效性/支出＝经济生产率。

如果预期目标是打印几份报告，那么，用可用的打印的报告数除以消耗资源（如工时、过程开销和材料损耗）来规定或测量生产率。

E.2.3.4 安全性

安全性测度与超时运行软件产品所产生的风险、使用条件及使用周境相关。安全性根据运行安全和应急安全来分析。运行安全是指软件在对环境和资源不产生危害的正常运行下，软件满足用户需求的能力。应急安全是指超出正常运行以外软件的运行能力以及改变资源阻止风险逐步扩大的能力。

E.2.3.5 满意度

满意度测度是指用户对产品的舒适程度及对产品使用的看法。

满意度可通过对标度的主观评级来规定和测量：如对产品的喜欢程度、对产品使用的满意度、在完成不同任务时可接受的工作负荷或特定使用质量目标的（如生产率、易学性）满足程度。其他满意度测

度可能还包括在软件使用过程中记录的正面和负面的评论。其他附加信息可以从长期测量中获得，如：缺席率，用户心理认为的或是实际体力所承受的工作的超负荷和欠负荷、健康问题报告或用户要求调换工作的频率。

满意度的主观测度是通过对用户的主观反映、态度或意见进行量化而获得的。量化过程可以有很多方式，如通过询问用户在任何特殊时刻的感觉的程度来量化，或按优先顺序评级，或根据问卷的势态标度量化。

如果使用得当，势态标度具有使用快捷、可靠性强及不需要特别的技巧的优点。使用心理测量学技术势态问卷是人们所熟知的，并且可对可靠性和有效性进行量化估计，也可以抵制如伪装的、正面的、负面的偏见及社会愿望的因素；其结果也可以和由过去获得的反映所确立的标准进行比较。见 F.3 中基于计算机系统测量满意度的问卷。

E.2.4 确立评估准则

使用质量的测度准则值的选择依赖于设置准则的组织的需要和产品需求。使用质量的目标和主要的目标(如写一封信)或子目标相关(如查询和替换)。将使用质量目标集中在最重要的用户目标上意味着忽略了许多功能，但这可能是最实际的方法。为具体子目标设置使用质量目标允许在开发过程的早期阶段进行评价。

为用户组设置准则值时，对于个体(如所有用户在 10 min 内完成任务)或是用户所占百分比(如 90%的用户在 10 min 内完成)来说，可以将准则值设为平均值(如完成任务的平均时间不超过 10 min)。

设置准则时，宜仔细地为每个测量项考虑合适的权值。如基于错误设置准则，必须分配权值以反映不同错误类型的相对重要性。

E.2.5 测度的解释

因为使用质量特性的相对重要性依赖于使用周境及规定或评价使用质量的目的，所以测度的选择和组合并没有通用的规则。

在将任何使用质量的测量结果推广至其他使用周境(不同的用户、任务和环境)时，宜小心。如果使用质量的测度是短时期内获得的，可以不考虑对使用质量有重要影响的罕见事件的取值。如间歇性系统错误。

对于通用的产品，通常有必要规定或测量不同的具有代表性的周境中的使用质量，这里的代表性的周境是指可能的使用周境的子集及可能被执行的任务子集。在这些周境中，使用质量也可能不同。

E.3 设计评价

宜尽可能地在与产品使用环境相接近的条件下评价产品。下列内容是重要的：

- 用户是指使用产品的具有代表性的用户；
- 任务是指系统要完成的具有代表性的任务；
- 条件是指产品使用的具有代表性的正常条件(包括获取帮助的方式，时间压力和干扰)。经验表明，通过控制评价条件，利用 8 个参与者的样本能获得可靠的结果。(见 F.2.4.1)

E.4 执行评价

E.4.1 执行用户测试和数据采集

评估使用质量时，用户独立工作是重要的。只有在正常使用条件下时，用户才能获得帮助。通常，对用户所碰到的问题记录，以及在使用阶段结束时与用户讨论并澄清的问题，也测量其有效性、生产率和满意度。使用录像技术对评价过程录像，有利于细节分析和录像结果剪辑。

如果通过远程监控用户，用户可以不受干扰的工作。

E.4.2 生成报告

如果需要综合的报告，通用行业格式(附录 F)提供了一种报告使用质量的良好结构。

附 录 F
（资料性附录）
使用质量测试报告的通用行业格式[1)]

F.1 目的和目标

易用性(usability)测试报告的通用行业格式(CIF)的主要目的是促使易用性结合为交互产品决策过程的一部分。这种决策过程的例子包括：购买、升级和自动化。它为供方的人类工效学工程师和易用性专业人员提供了一个向客户组织报告易用性测试的方法和结果的通用行业格式。

F.1.1 读者

供方组织中的易用性专业人员可利用 CIF 生成可供客户组织使用的报告。CIF 也供客户组织用来验证一份特定的报告是否符合 CIF。

易用性测试报告本身针对两类读者：

1) 客户组织的评价易用性测试技术价值和产品易用性的人类工效学工程师或其他易用性专业人员；

2) 其他专业技术人员和根据测试结果作出商业决策的管理者。

方法和结果部分针对第一类读者。这里对可重复测试方法和结果进行了技术细节描述。也支持对预期的开销和受益质疑的测试数据应用。为更好地使用它，需要有人类工程学或易用性工程的技术背景来理解和解释这些部分。引言针对第二类对象，它为非易用性专业人员和管理者提供概括信息。其他计算机专业人员一般也会对这部分感兴趣。

F.1.2 范围

CIF 报告格式的试验使用将在一个试验研究中产生。更多的关于试验研究信息请留意下列网站上的文档(http://www.nist.gov/usr/documents/whitepaper)。报告的格式假定在测试的设计和执行中已经遵循了合理的习惯(如参考文献[8,9])。推荐使用求和类型易用性测试。格式支持对任何经验的测试结果均有清晰和完全的报告。应使用可产生概括易用性测度的测试规程。对于一些易用性评价方法，如格式化测试：旨在确定问题而不是产生测度，目前的格式并不是针对这些测试方法的结果而构造的。此一般格式涵盖了需要报告的最少的信息。供方可以选择包含更多的信息。尽管可以将格式扩充到诸如具有用户接口的硬件等更广泛的应用，然而这次没有包括在内。当我们在试验性研究中获得更多的经验时这些问题将会被提出来。

F.1.3 与现行标准的关系

本文档并不是制订标准的正式参考文档，但现行的标准如 ISO 13407，ISO 9241-11 和 GB/T 18905.5—2002的附录 C 中已提供了资料。本文档和上述标准文档的主要部分是一致的，但在范围上有更多的限制。

F.2 报告格式描述

格式宜作为通用模板。根据客户组织、产品供方和任何第三方测试组织间所使用的协议来撰写所有部分的报告。

---

1) 附录 F 和 G 是由 IUSR 行业组织提供的，并非 ISO 的版权，在这里，是作为使用质量的测试结果报告的推荐用例。其最终版已作为美国国家标准 ANSI/INCITS-354—2001《易用性测试报告通用行业格式》而出版。出版这些附录中使用的术语"易用性(usability)"和 ISO 9241-11 中的使用质量的定义是相似的。(但不包括安全性，也不使用术语：生产率有效性)。

CIF 的组成部分或是强制的或是推荐的，并在正文中分别用记号“ ⁘”“◆”表示。

F.2.1 标题页

本部分包含：

⁘ 标明该报告作为通用行业格式文档(CIF)；声明 CIF 版本；

⁘ 命名测试的产品和版本；

⁘ 测试负责人；

⁘ 测试时间；

⁘ 报告起草日期；

⁘ 报告起草人；

⁘ 对测试方面所有问题都清楚的个人联系信息如电话、email、邮寄地址等，以便支持确认和复测。

F.2.2 执行概要

本部分提供测试的高层次概述。本部分以新一页开始，以一个分页符结束以便于将它用作一个独立的概要。本部分主要为客户组织提供决策方法的信息。这些人或许没有阅读本文档的技术内容，但是仍然会对下列问题感兴趣：

⁘ 产品的描述和识别；

⁘ 包括参与者及其任务的类型、数量的测试方法概要；

⁘ 以平均记分或其他合适的重要趋势的测度来表述结果；

◆ 测试的理由和性质；

◆ 执行结果的表格汇总。

如果声明的产品之间或数值之间存在差异，应说明这个差异不是偶然发生的概率。

F.2.3 引言

F.2.3.1 整个产品的描述

⁘ 本部分确定正式产品名称以及发布号或版本号。它描述此产品需要评价的部分。本部分还要确定：

⁘ 本产品的意向用户数；

◆ 任何具有特殊需求的组；

◆ 产品使用环境的简要描述；

◆ 产品支持的用户工作的类型。

F.2.3.2 测试目标

⁘ 本部分描述测试和任何特定兴趣领域的所有目标。可能的目标包括测试用户执行工作任务的情况及使用产品的主观满意度。本部分应包括：

⁘ 在测试中与用户直接和间接交互的产品的功能和部件；

◆ 如果被测产品的功能和组件是全部产品的子集，要解释集中在子集的原因。

F.2.4 方法

这是首要的关键技术部分，它必须提供充分的信息允许一个独立的测试者重复执行测试过程。

F.2.4.1 参与者

本部分根据人口统计学、专业经验、计算机经验和特殊需求来描述参加测试的用户。这些描述应该充分提供了通过对参与者类似取样就可以重复研究的信息。如果用户数总体和参与者样本之间有任何已知的差别，则应注意这种差别。如实际用户可能参加了培训而测试对象没有培训过。参与者不应来自同一个测试组织或供方组织。当报告人口统计组之间在易用性度量上的差别时，应给予特别注意。

通用描述应包括下列重要因素：

⁘ 参加受测人数，建议每个分组最少是 8 个人[10]。

- ⁘ 受测用户组的划分(如果超过一个受测用户组),如新手和专家。
- ⁘ 预期的要评价的用户组的能力和关键特性。
- ⁘ 选择参与者的方法及参与者是否有基本的能力和特性。
- ◆ 参与者样本是否包括具有特殊要求的代表性组。如年轻的、年老的、或智力和体力有缺陷的。

对于规定的测试参与者的特征和能力的表格,行宜表示每个参与者的情况,列宜表示每个特性的情况。宜选择与产品易用性相关的特性;宜允许客户来决定参与者与客户群的相似程度;并且参与者必须足够完整,这样可以保证召集到相似的参与者群。下表是一个例子;表中的特性是典型的但不可能覆盖所有的测试情况类型。

表 F.1

	性别	年龄	教育情况	职业/职责	专业经验	计算机经验	产品经验
P1							
P2							
Pn							

"性别"栏,显示男性或女性。

"年龄"栏,描述参与者的年龄,或在准确年龄未知的情况下填写年龄段(如 25～45)或年龄段分类(如 18 岁以下,65 岁以上)。

"教育情况"栏,描述受正规教育的年数(如在美国,受过中学教育就是 12 年,受过大学教育则是 16 年)。

"职业/职责"栏,描述当用户使用该产品时,他的工作角色。如果角色已知则填写角色名称。

"专业经验"栏,给出用户从事该角色的时间。

"计算机经验"栏,描述相关背景,诸如用户使用平台或操作系统,或产品范围等的经验。可能不止一列。

"产品经验"表示以前使用该产品或同类产品的时间和任何经验。

F.2.4.2 测试中产品的使用周境

这部分描述测试执行的任务、场景和条件,其中任务、应用系统的运行平台以及由测试参与者所做的特定配置是评价的一部分。评价的使用周境和预期的使用周境之间的任何差异都应在相对应部分注明。

任务

对参与者所执行的任务的充分描述对测试的正面有效性是关键的。

- ⁘ 描述测试的任务场景;
- ⁘ 解释选择这些任务原因(如频次最高的任务、最棘手的任务);
- ⁘ 描述任务的来源(如使用类似产品的客户的观察报告,产品市场规范);
- ⁘ 也包括任何提供给参与者的任务数据;
- ⁘ 为每个任务确立的任何完成的或是执行的标准。

测试设备

这部分涉及测试设备的物理描述。

- ◆ 描述进行评价的环境、空间(如实验室的易用性,方形的办公室,会客室、家庭办公室、寓所、场地);
- ◆ 详细说明任何影响结果的质量的相关特征和环境,如视频,音频录制设备、单向镜像和自动的数据采集设备。

参与者的计算环境⁘

这部分应包括确认测试和重复测试所要求的所有细节。包括参与者的计算机的合适的配置，如硬件型号，操作系统，版本和任何需要的环境和程序库。如果产品使用 WEB 浏览器，应标识浏览器的版本、名称及任何相关插件的版本。

显示设备⁘ 如果产品是基于显示器的可视化接口，则应该详细标明显示屏尺寸，监视器的分辨率及颜色设置(色素数)。如果是基于打印的可视化接口，则应该标明介质大小、打印机的分辨率。如果可视化接口的组件在尺寸大小上有不同，应明确测试中所使用的尺寸。这和字体有关。

音频设备◆ 如果产品有音频接口，要确定相关环境或音频位值，音量大小等。

手工输入设备◆ 如果需要使用手工设备，如键盘、鼠标、操纵杆等，要明确测试中所使用的设备结构和型号。

测试管理工具

⁘ 如果使用标准的问卷，可以在测试管理工具中加以明确描述。包括后附的定制问卷。

◆ 描述用于控制测试或记录数据的任何软件和硬件。

F.2.4.3 实验设计

⁘ 描述测试的逻辑设计，定义独立变量和控制变量。简短地描述源自于每种条件下记录的数据的测度。

规程

这部分细化了测试协议。

⁘ 给出测度的操作定义及任何存在的独立变量和控制变量。描述任何任务的时限，任何策略和培训、帮助、干扰，或对问题响应的过程；

◆ 包括从对参与者问候致意到解散的事件序列；

◆ 包括协议的机密性、完成形式、准备工作、预先培训和听取报告等细节；

◆ 查证参与者的学识，理解作为主体的人的权利和义务[1]；

◆ 规定评价团队遵循测试会话执行和数据记录的步骤；

◆ 规定在测试会话期间的参与者交互人数，简短描述他们所承担的角色；

◆ 声明出现在测试环境中的其他个体和他们的角色；

◆ 声明是否支付或赔偿参与者。

对参与者的一般性指导

⁘ 这里和附录中的所有指导都是为参与者提供的(参与者任务指导部分中提供的实际任务指导除外)；

⁘ 包括参与者在测试环境中与任何其他在场的人怎样进行交互的指导，包括如何请求帮助及如何与其他参与者交流。

参与者任务指导

⁘ 这部分是对任务指导的总结，附录中给出了精确的指导。

F.2.4.4 易用性度量

⁘ 解释所用的每类易用性度量的测度：有效性、效率和满意度。下面给出了这些度量的概念描述和实例：

有效性

有效性将使用该产品的目标与这些目标可达到的准确性和完备性相关联。一般的有效性测度包括任务完成的百分比、出错频率、协助参与者频率及在执行任务期间参与者获得文档或帮助的频率。该测度不考虑目标怎样达到，只考虑目标达到的程度。效率将有效性等级与资源开销的数量相关联。

完成率

结果应该包括参与者完全并正确地达到每个任务目标。如果目标可以部分完成(如未完成或未达

到最佳结果),则根据与部分结果值相关的特定标准按0～100%记分,得到所达到的平均目标。如拼写校对任务涉及确定和校正十个拼写错误,则完成率可以根据校正错误的百分率计算。另一个计算完成率的方法是权值:如在文档标题页中的拼写错误比其他页中的错误严重两倍。如果在报告中包括这种分析结果,就要说明进行分析时所选择的特定方法的基本原理。

注:应有独立完成率(如无测试者干预的完成率)和非独立完成率(有测试者干预的完成率)这两个不同度量的报告。

错误

错误是指测试参与者未成功完成的任务,或不止一次尝试完成部分任务。推荐使用数据记分法,包括根据某种分类法对错误进行分类的方法[2]。

帮助

当参与者不能继续一个任务时,有时测试管理者会直接给予程序上的帮助,以使测试继续进行。这种测试者的介入类型称作对该报告目的的"协助(assist)"。如果有必要给参与者提供协助,则必须确定在有协助和无协助条件下的效率和有效性度量。例如,如果参与者在任务A中接受了协助,在计算独立完成率时,就不应该包括任务A。不过,如果计算非独立完成率,则可以包括任务A。当允许或提供协助时,必须在测试结果报告中说明协助类型和次数。

在一些易用性测试中,当参与者不能完成任务时,他们可以得到类似于在线帮助或文档的指导。对本报告来说,获取相关的产品信息特征和在线帮助(help)不算是一种协助(assist)。不过,如果产品特征可以使参与者获得独立使用产品的能力,那么还是应该在报告中记录获取不同产品信息的次数。

效率

效率将所达到有效性级别的与资源开销的数量相关联。效率通常是用完成任务的平均时间来评估的。效率和其他资源如使用的总开销相关。一般效率的测度是用完成任务的时间来衡量的。

任务时间

结果必须包括所有参与者完成每个任务的平均时间,及时间范围和时间偏差标准。有时,需要更细的分类,如用户寻求或获得帮助的时间(包括文档、系统帮助或帮助桌面)。该时间应包括在任务的总时间里。

完成率/任务时间均值

完成率/任务时间均值测度是效率测度的核心测度。它确定在单位时间内顺利完成任务的用户的百分比(或目标完成的百分比)。公式表明随着时间增加,期望用户将会更成功。高效率产品意味着在较少的时间内用户顺利完成任务的百分比越高。它允许客户将快速但容易出错的界面(如有通配符命令行中删除文件命令)和慢但易使用的界面(如可以利用鼠标、键盘将文件拖至回收站)进行对比。

注:在特定使用周境中,即使在很难于解释的情况下,也要对有效性和效率的结果进行报告。必须要确定供方不考虑度量意义的原因。如假定产品的使用周境包括实时,及密切相关事务间的自由交互的环境,在这种情况下任务时间解释为效率测度或许没有意义。因为对于大多数用户来说,完成任务时间就是有效地使用时间。

满意度

满意度是指使用产品时,用户的主观反应。用户满意度可能和使用产品的动机相关,可能影响某些情况下的使用性能。用于测量满意度和相关看法的问卷,通常使用Likert标度和语义分化标度来构造。

可以使用不同的仪器测量交互软件产品用户满意度,许多公司自行建立了自己的仪器。不管是使用外部的、标准的或定制的仪器,建议包含主观评级方法如满意度,有用性,和易使用性,因为客户组织普遍对此感兴趣。

许多问卷方法被广泛使用,包括ASQ[5],CSUI[6],PSSUQ[6],QUIS[3],SUMI[4],SUS[7]。尽管每种问卷法给出了产品易用性主观测度,但大多数都包括满意度、有用性、易使用性。

供方可以选择使用已经确认的、发布的满意度测度或提交他们自己开发的度量。

结果

这是报告中的第二个主要的技术部分。包括对数据记分、简化和分析方法的说明，给出了定量化格式方面的主要研究成果。

数据分析

数据记分⁘

应该对采集的数据记分方法作详细的描述，这样以便于在其他组织重复测试时，可以使用重复记分方法。尤其是对排除在外的人员的范围、错误数据分类和独立完成/非独立完成的记分等特定项进行详细描述。

数据简化⁘

应该对采集的数据简化方法作详细的描述，这样以便于在其他组织重复测试时，可以使用数据简化方法。尤其是对任务和任务分类中的数据崩溃等特定项要详细描述。

统计分析⁘

应该对数据分析方法作详细的描述，这样以便于在其他组织重复测试时，可以重复使用数据分析方法。尤其是对统计过程(如数据转化)和检验(t-检验，F 检验和区间的显著性检验)要详细描述。报告中的平均记分必须包括标准偏差估计和可选的均值的标准误差。

结果表述

⁘ 必须对有效性，效率和满意度测量的结果进行报告。

应该包括图表形式的两种结果在内。偶尔使用不同的图形格式描述易用性数据是有效的。附录 C 中有测试报告的样例。条形图用来表示主观数据如利用 likert 标度方法采集的数据。可以通过有效使用不同的图形对专家基准时间和参与者执行的平均时间进行对比。可以对这些数据结果有一些简短的解释，但是不必详细地描述。

执行结果⁘

建议对参与者以每个单位任务为基础将有效性和效率的结果制成表格。对于相关任务的组来说(一个组包括所有的程序编制任务，另一组包括调试任务)，表格的结果和表述是合理的。如果单元任务有子任务，那么在总结中记录子任务。如，如果单元任务是明确一页中的拼写错误单词，那么可以对发现错误拼写的百分率汇总，最终，汇总表中显示总的平均任务时间和完成率。如果和产品设计特殊应用相关，测试者应该附上另外的度量表。

表 F.2 任务 A

用户 #	无协助情况下的任务有效性[(%)完成量]	有协助情况下的任务有效性[(%)完成量]	任务时间分	有效性/任务时间均值	错误	协助
1						
2						
N						
均值						
标准偏差						
最小值						
最大值						

表 F.3 汇总

用户 #	无协助情况下的任务有效性总量[(%)完成量]	有协助情况下的任务有效性总量[(%)完成量]	总任务时间分	有效性/任务平均时间	总错误	总协助
1						
2						
N						
均值						
标准偏差						
最小值						
最大值						

满意度结果❖

满意度问卷的数据可以采用与上述执行数据相同的汇总方法。每列表示一个单独的测量标度。

表 F.4 满意度

用户 #	标度 1	标度 2	标度 3	…	标度 *N*
1					
2					
N					
均值					
标准偏差					
最小值					
最大值					

F.2.5 附录

定制问卷、参与者一般指导和参与者任务指导可作为附录。发布的注释也应作为单独的附录,它包含供应商希望包含的自测试开始以来的可以解释或更新的测试结果的任何信息(例如,自测试以来,UI(用户界面)设计已确定的情况)。

F.3 参考文献

[1] American Psychological Association. Ethical Principles in the Conduct of Research with Human Participants. 1982.

[2] Norman, D. A. (1983) Design Rules Based on Analyses of Human Error. Communications of the ACM, 26(4), 254-218.

[3] Chin, J. P., Diehl, V. A., and Norman, K. (1988). Development of an instrument measuring user satisfaction of the human-computer interface. In the Proceeding of ACM CHI '88 (Washinton D. C.), 213-218.

[4] Kirakowski, J. (1996). The software usability measurement inventory: Background and usage. In Jordan, P., Thomas, B., and Weerdmeester, B. (Eds.), Usability Evaluation in Industry. UK: Taylor and Francis.

[5] Lewis, J. R. (1991). Psychometric Evaluation of an After-Scenario Questionnaires for Computer Usability Studies: the ASQ. SIGCHI Bulletin, 23(1), 78-81.

[6] Lewis, J. R. (1995). IBM Computer Usability Satisfaction Questionnaires: Psychometric Evaluation and Instructions for Use. International Journal of Human-Computer Interaction, 7, 57-78.

[7] Brooke, J. (1996). SUS: A "quick and dirty" usability scale. Usability Evaluation in Industry. UK: Taylor and Francis. (http://www.usability.serco.com/trump/documents/Suschapt.doc).

[8] Rubin, J. (1994) Handbook of Usability Testing, How to Plan, Design, and Conduct Effective Tests. New York: John Wiley & Sons, Inc.

[9] Dumas, J. & Redish, G. (1993). A Practical Guide to Usability Testing. New Jersey: Ablex Publishing Corp.

[10] Nielsen, J. & Landauer, T. K. (1993) A mathematical model of the finding of usability problems. In :CHI '93. Conference proceedings on Human factors in computing systems, 206-213

附 录 G
（资料性附录）
通用行业格式易用性测试实例[2)]

DiaryMate 版本 1.1

报告者：A Brown

C Davidson

Super Software Inc

（1999 年 9 月 1 日）

1999 年 8 月测试

如果对该报告有疑问，请致信至：
E Frost，易用性经理
Super software Inc
19483 Outerbelt Ave
Hayden CA 95014 USA
408 555-2340
Efrost@supersoft.com

2) 附录 F 和 G 是由 IUSR 行业组织提供的，并非 ISO 的版权，在这里，是作为使用质量的测试结果报告的推荐用例。这些附件中所使用的术语“易用性”和 ISO 9241-11 中的使用质量的定义是相似的。（但不包括安全性，不使用生产率的效率术语），附录 G 是根据真实的评价编写的样例。

G.1 引言

G.1.1 执行概要

DiaryMate 软件是计算机版的日记和通讯录。DiaryMate 软件为个人和工作组提供了日记、联系和会议管理的功能。该测试是证明 DiaryMate 软件的安装、日历表示和通讯录任务对秘书和经理的易用性。

给 8 个经理提供磁盘和用户手册，并要求他们安装该产品。在已经花费了一定时间熟悉软件后，要求他们使用该软件的功能添加新的联系信息和安排会议。

所有参与者在平均 5.6 min 安装产品(尽管很小的安装子组件遗失了)。所有的参与者成功的添加上联系信息。完成的平均时间是 4.3 min。

8 个人中有 7 个人在 4.5 min 内安排了会议。

SUMI 问卷总得分为 51。对于所有的标度，目标值 50 是在 95%置信度内(行业平均 SUMI 记分)。

G.1.2 整个产品描述

DiaryMate 软件是计算机版的日记和通讯录。DiaryMate 软件为个人和工作组提供了日志、联系和会议管理的功能。这是一个包括在线帮助和 50 页的用户手册商业产品。

主要用户组是办公室工作人员，具有代表性的初级和中级的经理。DiaryMate 软件要求 Windows 3.0 以上，用户应该有基本的 Windows 知识。在网站 www.supersoft.com/diarymate 上提供了所有的技术规范。

G.1.3 测试目标

评价目的是确认日历和通讯录功能的易用性，这是 DiaryMate 软件的主要特征。要求具有代表性的用户完成典型任务，并采用有效性、效率、满意度测度。

预期安装少于 10 min。所有的用户在 5 min 内添加联系信息。SUMI 总得分应该比行业平均分 50 还要高。

G.2 方法

G.2.1 参与者

打算使用的周境：期望的 DiaryMate 软件用户的关键特征和能力是：

- 熟悉 PC 和基本的微软 Windows 知识
- 英语运用能力
- 精通办公室业务

每天至少花 10 min 处理日记和联系信息的相关任务。

用户还应该有一些特征可能影响 DiaryMate 的易用性：

有一定的使用微软 Windows 的经验。

- 有使用其他日记应用软件的一定经验
- 对使用电子日记任务软件持有的态度
- 当前工作中的工作职能和从事该项工作的时间

测试使用的周境：选择 8 个具有关键特征和能力的初级或中级经理，他们以前并没有使用过 DiaryMate 软件。对其他可能影响易用性的参与者的特征以及年龄段、性别都进行了记录。

	职务	工作时间/年	使用 Windows 的经验/年	使用电子日记的经验/年	对电子日记的态度 (1～7)[a]	性别	年龄组
1	中级经理	5.5	3.5	0.0	6	女	20～35
2	初级经理	0.8	2.1	0.8	1	女	20～35
3	中级经理	2.1	2.5	2.1	3	男	20～35
4	初级经理	4.9	3.5	1.5	2	女	36～50
5	中级经理	0.7	0.7	0.7	2	男	20～35
6	初级经理	1.6	2.1	0.0	3	女	36～50
7	中级经理	4.3	1.4	0.0	4	男	36～50
8	初级经理	2.7	4.6	2.7	4	男	20～35

[a] 1＝喜欢尽可能多用计算机，7＝宁愿尽可能少用计算机。

G.2.2 测试中产品使用的周境

G.2.2.1 任务

打算使用的周境：通过与潜在用户的沟通使其明确软件安装是一项重要的任务。在对软件熟练应用后，其他的关键任务就是添加联系信息，以及安排一次会议。

测试使用的周境：为了评价所选择的任务是：

- 为参与者提供应用程序的拷贝和文档，并要求其安装该软件；
- 每个参与者重新启用程序，并花一定的时间熟悉 DiaryMate 软件的日记和通讯录功能；
- 要求参与者利用信息支持功能添加新的联系信息；
- 要求参与者使用日记功能安排会议。

G.2.2.2 测试设备

打算使用的周境：办公环境。

测试使用的周境：评价在 Hayden 的易用性实验室完成。测试室是封闭的，配有桌子、椅子和其他办公家具。参与者不受干扰。可以通过单向镜像，及电视录像和远程屏幕观察参与者。

G.2.2.3 参与者的计算环境

打算使用的周境：DiaryMate 软件运行在奔腾处理器，操作系统为 Windows，至少 8 M 内存的 PC 机上。

测试使用的周境：标准的配置为 Netex PC-560/1(pentium 60，32 M 内存)，Netex 鼠标，432 mm (17 英寸)分辨率为 800×600 的显示器，操作系统为 Windows 95.

G.2.2.4 测试管理工具

使用 Hanks 软件进行易用性登录计时。尽管电视录像的信息不作为报告的一部分,但仍使用电视录像录制会话(参与者视图和屏幕图形结合起来)。会话结束时,参与者完成主观评级标度及 SUMI 满意度问卷。SUMI 记分为均值 50 和标准偏差 10(根据对欧洲和美国的被测的 200 个办公类型标准样本,详细信息参见,www.ucc.ie/hfrg/questionnaires/sumi/index.htm)。

G.2.3 测试设计

对 8 个初级和中级的经理进行测试。

对于以下 3 种任务,可以计算完成率的平均值,完成目标的平均值、平均任务时间,平均完成率的效率和完成目标的效率。

- 安装产品;
- 添加新的联系信息;
- 安排会议。

G.2.3.1 规程

当参与者到达时,告诉他们正在测试 DiaryMate 软件的易用性,目的是为了了解 DiaryMate 软件是否符合用户要求,而不是对他们能力的测试;将参与者带至评价地点,包括控制室,告知参与者要对整个测试过程进行录像;要求参与者签一个发布形式,并要求他们确认在参加测试前已提供的信息:工作描述,工作时间(年),使用 Windows 的经验(年),使用电子日记的经验(年)及年龄组。他们自己也要对支持日记和联系管理任务的计算机应用的态度评分(等级为 1~7),如喜欢尽可能多用计算机、宁愿尽可能少用计算机。

先给参与者入门指导,评价者在每个任务开始前重置计算机,并提供下一个任务的指导。并告诉参与者为每个任务所分配的时间。要求参与者完成任务后就告诉评价者(通过电话)。不给参与者提供任何外部帮助。

在最后一个任务完成后,要求参与者完成主观评级及 SUMI 的问卷。

评价者询问参与者所遭遇的困难。(这个信息不在报告中出现)

最后付给参与者 $75。

G.2.4 度量

G.2.4.1 有效性

完成率:正确完成每个任务的参与者的百分比;

目标完成均值:是指正确完成每个任务的程度的均值,以百分比记分;

差错:不测量差错;

协助:不给参与者提供任何协助。

G.2.4.2 效率

任务时间:正确完成每个任务的平均时间(对于已完成的正确任务来说);

完成率效率:完成率均值/任务时间均值;

目标完成率:完成目标均值/任务时间均值;

查阅手册次数:对手册单独进行查阅的次数。

G.2.4.3 满意度

满意度使用主观评级测量和 SUMI 问卷,在会话结束时,参与者对产品的总的满意度,效率,影响,可控制性和易学习性记分。

G.3 结果

G.3.1 数据处理

G.3.1.1 数据记分

目标完成均值:指正确完成每个任务的程度的均值,以百分比记分。

通过和几个用户讨论潜在的日记和联系信息差错对业务的影响,得出了下列计算目标完成的均值的记分机制。

- 安装:所有组件成功安装:100%;安装中每有一个被省略的必要子组件就扣除 20%。
- 新联系:所有内容输入正确:100%;每漏掉一项信息,扣除 50%;错误字段中的每项信息,扣除 20%;每个排版问题扣除 5%。
- 新会议:所有内容输入正确:100%;错误时间或数据:0%;错误字段中的每项信息,扣除 20%;每个排版问题扣除 5%。

综合所有扣除量,等于或超过 100%,那么完成的目标的记分为 0%。

G.3.1.2 数据简化

除了每项任务的数据,综合结果显示了有效性和效率度量的总任务时间和方法结果。

G.3.1.3 数据分析

使用 SUMI 记分程序(SUMISCO)分析 SUMI 的结果。

G.3.2 执行结果

SUMI 满意度问卷的分数是 51。目标值为 50(行业平均 SUMI 分数)说明在所有领域中达到 95% 的置信度。

G.3.2.1 安装

所有参与者在 5.6 min 的平均时间内成功安装产品(虽然漏掉了一个次要的子组件)。

参与者 #	没有协助情况下的任务完成率/%	目标完成率/%	任务时间/min	完成率/任务时间/%/min[a]	查阅手册次数
1	100	100	5.3	19	1
2	100	100	3.9	26	0
3	100	100	6.2	16	1
4	100	80	9.5	11	2
5	100	100	4.1	24	0
6	100	100	5.9	17	1
7	100	100	4.2	24	0
8	100	100	5.5	18	0
均值	100	98	5.6	19	0.6
均值的标准差	0.0	2.5	0.6	1.8	0.3
标准偏差	0.0	7.1	1.8	5.1	0.7
最小值	100	80	3.9	11	0.0
最大值	100	100	9.5	26	2.0

[a] 当在两个产品间进行比较时,这种每分钟完成的百分比的组合数字是有用的。通过完成目标/任务时间就得到了一种相关的测度方法。

G.3.2.2 添加新的联系信息

所有的参与者成功地增加了新的联系信息(其中两人有排版错误)。完成任务的平均时间是 4.3 min。

参与者#	没有协助情况下的任务完成率/%	目标完成率/%	任务时间/min	完成率/任务时间均值/(%/min)	查阅手册次数
1	100	100	4.4	23	0
2	100	100	3.5	29	0
3	100	95	4.6	22	1
4	100	100	5.5	18	1
5	100	100	3.8	26	0
6	100	100	4.5	22	0
7	100	95	4.9	20	1
8	100	100	3.3	30	0
均值	100	99	4.3	24	0.4
均值的标准差	0.0	0.8	0.3	1.5	0.2
标准偏差	0.0	2.3	0.7	4.2	0.5
最小值	100	95	3.3	18	0.0
最大值	100	100	5.5	30	1.0

G.3.2.3 安排会议

8个参与者中有7个在4.5min的平均时间内成功地安排了会议。有些信息没有填进预期的字段中,这些做了标记的字段在产品的发布版本中已经做了改进。

没能完成任务的参与者在此之前从未使用过电子日记,并对此持否定态度。菜单结构随后得到了改进,使之可以说明安排的过程。

参与者#	没有协助情况下的任务完成率/%	目标完成率/%	任务时间/min	完成率/任务时间均值/%/min	查阅手册次数
1	0.0	0.0	0.0	0.0	3
2	100	95	4.2	24	2
3	100	80	5.6	18	0
4	100	100	3.5	29	1
5	100	90	3.8	26	1
6	100	60	6.1	16	0
7	100	75	4.6	22	0
8	100	80	3.5	29	2
均值(#2~7)	100	73	4.5	22	1.1
均值的标准差	0.0	4.8	0.4	1.7	0.4
标准偏差	0.0	13.5	1.0	4.9	1.1
最小值(#2~7)	100	60	3.5	16	0
最大值(#2~7)	100	100	6.1	29	3

注:已给出7个参与者完成任务的汇总数据。

G.3.3 综合执行结果

参与者#	没有协助情况下的任务完成率/%(全部任务)	平均目标/(%)	总任务时间/min	完成率/总任务时间/(%/min)	查阅手册次数
1	67	67	9.7	7	4.0
2	100	98	11.6	9	2.0
3	100	92	16.4	6	2.0
4	100	93	18.5	5	4.0
5	100	97	11.7	9	1.0
6	100	87	16.5	6	1.0
7	100	90	13.7	7	1.0
8	100	93	12.3	8	2.0
均值(#2～7)	100	93	14.4	7	1.9
均值的标准差	0.0	1.5	1.0	0.5	0.4
标准偏差	0.0	3.9	2.7	1.3	1.1
最小值(#2～7)	100	87	11.6	5	1.0
最大值(#2～7)	100	98	18.5	9	4.0

注：已给出7个参与者完成任务的汇总数据。

G.3.4 满意度结果

G.3.4.1 主观评级结果

根据“7点两端(7-point bipolar)”的Likert类型标度得到主观评级的数据，1＝最坏等级，7＝最好等级，关于不同标度的评级如下：

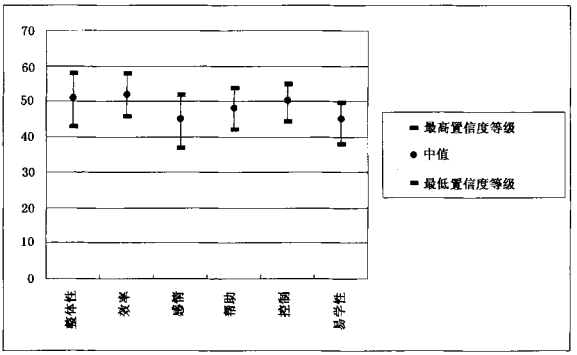
参与者#	满意度	有用性	易用性	清晰度	吸引性
1	5	3	3	3	4
2	5	6	6	5	5
3	5	5	4	5	6
4	2	5	4	2	5
5	4	4	4	4	5
6	4	4	6	5	6
7	3	2	4	2	3
8	6	6	4	5	6
均值	4.3	4.4	4.4	3.9	5.0
标准偏差	1.3	1.4	1.1	1.4	1.1
最小值	2	2	3	2	3
最大值	6	6	6	5	6

G.3.4.2 SUMI结果

SUMI满意度问卷的分数是51。目标值为50(行业平均SUMI分数)说明在所有标度中达到95%的置信度。

参与者#	整体性	效率	感情	帮助	控制	易学性
1	35	39	33	30	40	42
2	50	62	33	44	54	36
3	55	52	45	53	46	49
4	51	53	51	52	55	47
5	48	45	44	46	48	42
6	51	59	36	45	53	38
7	54	52	46	52	47	50
8	52	49	49	53	56	48
中值	51	52	44	49	50	44
置信度上限	58	58	51	55	56	50
置信度下限	44	46	37	43	44	38
最小值	35	39	33	30	40	36
最大值	55	62	51	53	56	50

整体性测度给出满意度的总体情况。效率表示的是参与者对软件效率的感知，感情表示他们对该产品的喜欢程度，帮助表示参与者可以找到软件帮助的程度，控制表示他们在控制方面的感受，易学性是参与者对软件是否易于学习的感觉。



G.4 附录A 对参与者的指导

G.4.1 对参与者的一般性的指导

谢谢您在这次评价中给予的帮助。

这次测试的目的是使人们认识到使用 DiaryMate，及其日记和联系的管理功能是如何方便。

为了完成这个目的，我们将要求您执行一些任务，您的执行情况将被录像并分析。为了有助于我们理解结果，我们将请求您完成标准问卷并回答几个关于您自己的和您的工作地点的问题。

评价的目的就是帮助评估产品，而且结果也有助于新版本的设计。

请记住，我们测试的是软件，而不是您。

当您完成每个任务时，请打电话-1234 告诉我们。

在此期间，我们不能给予您任何协助。

G.4.2 参与者任务说明

您已经收到了 DiaryMate 软件的拷贝，在您看到产品以前，您非常渴望了解产品以便于知道是否符合您的业务要求。

您将执行以下任务：

1. 安装软件
2. 重启程序，并花一定的时间熟悉软件的日记和通讯录的功能
3. 利用提供的信息，在通讯录中添加一个详细的新的联系信息
4. 使用日记功能安排会议

我们感兴趣的是您是如何使用 DiaryMate 软件完成这些任务的，及您觉得该软件是否有用。

如果您已经准备开始，请告诉我们！

任务 1——安装软件

（限 15 min 内装好）

桌上有一个标有 DiaryMat 的信封，内装有 DiaryMate 软件磁盘、指导手册。

如果准备好了，请安装软件。

所有的您需要的信息均在信封里。

如果要继续执行下一个任务，请告诉我们！

任务 2——熟悉期

花一些时间熟悉软件的日记和通讯录的功能。

（不能超过 20 min）

如果要继续执行下一个任务，请告诉我们！

任务 3——添加联系信息

（不超过 15 min）

使用软件添加如下的信息

姓名：Dr Gianfranco Zola

公司：chelsea Dreams Ltd

地址：25 Main street

los angeles

California 90024

电话（工作室）222 976 3987

（宅）222 923 2346

如果要继续执行下一个任务，请告诉我们

任务 4 ——安排会议

（不超过 15min）

使用软件安排会议：

日期：2001 年，11 月 23 日

地点：The Blue Flag Inn. cambridge

时间：中午 12:00～下午 1:30

出席者：参与者本人和 Dr Gianfranco Zola

如果已完成任务，请告诉我们！