文章编号: 1006-8309(2000)04-0035-04

# 诊查型用户界面可用性评价方法(IM)(下)——比较与建议

吴昌旭, 张侃
(中国科学院心理研究所, 北京 100101)

摘要: 首先对文(上)评述的 7 种诊查型方法(Inspection Methods, IM)进行了多维度定性综合比较, 使评价人员能在实际运用中, 根据不同方法的特点与要求选择合适的评价方法; 通过比较还发现其中的工效学检查表有较多优势; 然后, 本文对各种诊查型方法之间进行定量比较的方法作了探新, 并在总体上对现有诊查型方法存在的问题提出了一些建议。

关键词: 人机界面; 用户界面; 可用性评价; 诊查型方法
中图分类号: TB18; TP31 文献标识码: A

## 1 诊查型各方法综合比较

7 种诊查型方法虽各自均有存在优缺点, 但它们之间缺少系统的比较。没有比较, 评价人员就很难在这些方法中依据实际情况选择合适的方法; 没有比较, 就很难分析方法之间的功效差异。

### 1.1 各诊查型方法之间的多维度定性比较

我们综合了 Dix, Virzi 与 Neilsen 等的研究结果[1,2,3], 运用了 Barry Kirwan 比较多种人误预测法的比较方法[4], 对这 7 种方法进行了多维度定性比较, 见表 1。

比较结果: 能使人机界面评价人员根据不同方法的特点与要求, 选择较为合适的可用性评价方法; 通过计分计算发现, 在无需可用性专家参与的各方法中, 工效学检查表(累计得分 25.5)具有较多的优点与较高的可操作性, 尤其适合在我国目前缺少可行性评价专家等情况限制下运用。

### 1.2 各诊查型方法之间的定量比较的方法

从目前所收集到的资料看, 尚未见到有关这些方法之间的定量比较的研究, 因此我们认为有必要对定量比较的方法加以探索。

1.2.1 对方法的评分者间信度的计算 可以求两个或两组评价人员的所预测的可用性问题的类型与数量的相关系数, 或者以预测的可用性问题的严重等级计算不同方法的肯德尔和谐系数; 方法的评分者间信度越高, 说明这一方法获得的评价结果就越不容易受到评价人员个体差异的影响。

1.2.2 对方法同时效度(Concurrent Validity)的计算 借用对不同人因错误分析技术同时效度的计算方法:[6]

设 $a$ = 用 IM 预测到的和实际观察到的可用性问题一致的问题数量(即击中的问题数量)

---

关分析找出各种错误与事故之间的关系, 对于通过前馈控制预防各种失误和事故会有积极的作用。

## 参考文献

[1] Robert L, Helmreich. Managing human error in aviation[J]. Scientific American, 1997, 276(5): 62-68.
[2] Reason J. Human Error[M]. UK, Cambridge: Cambridge University Press, 1990.
[3] 朱祖祥. 人类工效学[M]. 杭州: 浙江教育出版社, [J].
[4] Reason J. Driving errors, driving violations and accident involvement[J]. Ergonomcs, 1995, 38(5): 1036-1048.
[5] 王二平. 人误研究的组织定向[J]. 人类工效学, 1999, 5(1): 44-47.
[6] Rasmussen J. Mental procedures in real-life tasks: A case study of electonic trouble shooting[J]. Ergonomics, 1974, (17): 293-307.
[7] 高佳, 黄祥瑞. 人的可靠性分析研究的进展[J]. 人类工效学, 1996, 2(4): 52-57.
[8] Fleishman ED, Buffardi LC, Morach RA, et al. Development of a Model to predict Human Error Rates From the Ability Requirement of Job Tasks[M]. NRG-04-91-361, 1994.

[收稿日期] 1999-12-13
[修回日期] 2000-10-27

$b$= 实际观察到的可用性问题数量;

$c$= 用 IM 预测到的"可用性问题",但未被实际观察到的问题数量(虚报的问题数量);

$d$= 未被 IM 预测到的,但被实际观察到的问题数量(漏报的问题数量);

$e$= 被 IM 预测到的的实际不可能出现的可用性问题数量(正确拒绝的问题数量);

方法的同时效度= $1/2\{a/b+[1-(c/e)]\}$

方法的同时效度越高,说明这一方法所预测的可用性问题与实际观察到的可用性问题的一致程度也越高。

1.2.3　对方法的辨别力与反应偏向的计算　在计算同时效度的基础上,我们提出了运用信号侦察论(SDT)来比较各方法的辨别力,见表 2。

表 1　诊查型评价方法的具体比较

| 比较的维度 | 认知走查法 | 启发式评价 | 合作性评价 | DEA | 工效学检查表 | 小组评价法 | 专家评价法 | 项目权重 |
|---|---|---|---|---|---|---|---|---|
| 目的与侧重 | 考察易学性 | 根据某些原则发现可用性问题 | 从用户报告发现可用性问题 | 预测任务完成中可能出现的交互错误 | 按项目逐条发现可用性问题 | 发现可用性问题及考察可用性是否达到标准 | 根据专家经验,发现可用性问题 | / |
| 适用的开发阶段 | 整个,主要是在早期 | 主要是在早期 | 整个 | 整个 | 整个 | 主要是在早期 | 主要是在早期 | / |
| 发现可用性问题的全面性(外部效度) | 低(1) | 高(3) | 低(1) | 低(1) | 高(3) | 高(3) | 高(3) | 2 |
| 方法构建的理论背景(构思效度) | 高(3) | 中(2) | 低(1) | 中(2) | 中(2) | 低(1) | 低(1) | 1 |
| 评价者经验影响程度(评分者间信度) | 低(3) | 高(1) | 高(1) | 低(3) | 低(3) | 高(1) | 高(1) | 1 |
| 是否必须是专家 | 否①⑤(3) | 否①(3) | 否(3) | 否(3) | 否(3) | 至少一名专家(2) | 是(1) | 1/2 |
| 评价者经验要求 | 高①(1) | 中②(2) | 低(3) | 中(2) | 低(3) | 高(1) | 高(1) | 1/2 |
| 评价者人数 | ⑥(2) | 至少 3 ~ 5 人⑥(1) | 2 人以上(2) | 1 人亦可(3) | 1 人亦可(3) | 2 人以上(2) | 至少 3 ~ 5 人⑥(1) | 1/2 |
| 使用真实用户人数 | 0(3) | 0(3) | 至少 1 人(1) | 0(3) | 0(3) | 0(3) | 0(3) | 1/2 |
| 对原形界面交互程度要求 | 低(3) | 低(3) | 高(1) | 低(3) | 低(3) | 低(3) | 低(3) | 1/2 |
| 评价时间耗用 | 多(1) | 少(3) | 少(3) | 少(3) | 少(3) | 少(3) | ④(2) | 1/2 |
| 是否对特定任务进行分析 | 是(3) | 否(1) | 是(3) | 是(3) | 可能(2) | 可能(2) | 可能(2) | 1/2 |
| 是否推论错误原因 | 是(3) | 可能(2) | 可能(2) | 是(3) | 否(1) | 可能(2) | 可能(2) | 1/2 |
| 是否推论用户心理过程与原因 | 是(3) | 否(1) | 否(1) | 可能(2) | 否(1) | 否(1) | 否(1) | 1/2 |
| 是否估计界面学习时间 | 可能(2) | 可能(2) | 否(1) | 否(1) | 否(1) | 可能(2) | 否(1) | 1/2 |
| 结果性质 | 定性(2) | 定性(2) | 定性(2) | 定量(3) | 定量(3) | 定性(2) | 定性(2) | 1/2 |
| 提供信息水平 | 细节(2) | 整体(2) | 以细节信息为主(2) | 细节(2) | 整体/细节(3) | 整体/细节③(2) | 整体(2) | 1/2 |
| 累计得分 | 22 | 21.5 | 16 | 21.5 | 25.5 | 20.5 | 18.5 | |
| 名次 | 2 | 3 | 7 | 3 | 1 | 5 | 6 | |

注:①增加评价者人数可以弥补专家的缺乏;

②增加评价者人数可以弥补评价经验的不足;

③随不同成员组成而定;

④随专家人数增加而减少;

⑤专家(特别是对系统与用户均熟悉的专家)的参与能大幅度提高方法运用的有效性;

⑥随评价人数的增加,该方法发现问题的能力增加;

( )内 1 - 3 分值与权重(以总分 10 分计)的确定是根据各方法的评述(见本文(上))[20]与评分准则[4](见表 3)及相关文献。[1,2,3,5,19]

表 2　信号侦察论的刺激 - 反应矩阵[7]

| $S$ | 做"是"反应 | 做"否"反应 |
|---|---|---|
| 信号 | 击中的问题数量 $a$ | 漏报的问题数量 $d$ |
| 噪声 | 虚报的问题数量 $c$ | 正确拒绝的问题数量 $e$ |

$$\left.\begin{aligned} P_{击中} &= a/(a+d) \Rightarrow Z_{击中},\ O_{击中} \\ P_{虚报} &= c/(c+e) \Rightarrow Z_{虚报},\ O_{虚报} \end{aligned}\right\} \Rightarrow \text{方法的}$$

辨别力 $d' = Z_{击中} - Z_{虚报}$　方法的反应偏向 $\beta = O_{击中}/O_{虚报}$

重要的是,运用 SDT 可以分离 IM 的辨别力

与反应偏向。$d'$ 越小, 说明方法在大量“噪声”中检出可用性问题的能力越差; β 越小, 说明使用该方法时, 更易虚报假的可用性问题。如果能进一步结合下面的“播种”问题方法, 改变“播种”问题出现的先定概率, 则可以求出不同 IM 的 ROC 曲线。

1.2.4 对 IM 预测可用性问题的绩效计算 引入比较多种预测软件出错量方法绩效的思路[8], 首先有意在界面中有意设计(“播种”)一定数量的可用性问题, 让不知道这些问题的评价人员使用某个可用性评价方法对这一界面进行评价, 可以得到该方法的评价效力。

方法的评价效力= 被该方法发现的“播种问题”的数量/总共的“播种问题”的数量

当然, 这里需要进一步进行各 IM 之间的定量比较的实验研究, 才能实证这些定量比较思路的有效性。

表 3 部分评分准则[4]

| 维度 | 高 | 中 | 低 |
|---|---|---|---|
| 发现可用性问题的全面性 | 从界面的可用性各个角度出发考察其可用性 | 从界面的可用性几个角度出发考察其可用性, 但不够全面 | 仅从界面的可用性一个角度出发考察其可用性 |
| 方法构建的理论背景 | 该方法的提出是基于一个系统的心理学理论 | 该方法遵循着一定的理论评价方法与原则, 但缺乏系统的理论支持 | 完全通过评价人员的主观经验及用户的意见, 没有一定的理论支持 |
| 受评价者经验影响程度 | 评价过程无章可循, 评价的经验直接决定方法运用的效果 | 评价过程能够依据一定的评价步骤进行, 但效果仍受到评价经验的影响 | 评价者只需依据方法的步骤要求进行分析预测, 经验对评价结果的影响较小 |
| 对评价者的经验要求 | 必须具备大量的人机系统界面可用性评价经验与工程心理学知识 | 须具备可用性评价方法知识与短期训练 | 无须特殊经验与知识 |

## 2 诊查型方法存在的问题、相应建议与发展趋势

### 2.1 方法运用的模糊性及其出路

纵览以上方法, 对于一般评价人员来讲, 他们仍难以依据非常严格的评价步骤有效地评价界面, 这使得评价人员的经验水平成为影响评价质量的关键。对于这一缺点, 我们建议:

(1) 广泛运用工效学检查表, 将可用性评价准则可操作化, 例如使用工效学检查表对目前发展最为迅速的网页设计进行评价[9]。

(2) 还可以采用基于任务的精细分析技术, 如近年来出现的 TAFEI(基于任务分析的人因错误预测)、PHEA 或 SHEARPA(人因错误预测分析方法)。[6,10]

依据这些技术, 再辅以计算机支持的实时评价系统,[11,12] 可使设计早期的可用性评价更为简易、迅速, 这也可能成为今后 IM 发展的新趋势。

### 2.2 难以替代可用性测试

#### 2.2.1 IM 相对于可用性测试的局限

IM 可以作为可用性测试的前驱, 但不能代替可用性测试;[2] 可用性测试比 IM 能发现更多的可用性问题,[13] 并且能直接考察用户绩效水平。[2] 此外, IM 还在以下几个方面显得捉襟见肘:[18]

(1) 预测整个界面的可接受性及相对于市场上同类产品的优越性;

(2) 从用户特殊环境出发, 权衡交互绩效的某些方面(如速度与准确性的权衡);

(3) 进行产品在功能、尺寸、成本方面的整体权衡;

(4) 直接从用户、销售商、购买商及产品的全球化角度评价整个界面。

#### 2.2.2 解决思路

IM 中的部分方法与可用性测试具有互补性。[14] 因此, IM 的局限可以由可用性测试弥补。有研究者提出了一种结合方式: 廉价可用性评价方法(DUE)。它主要是在创建交互画面情景(scenario)(整个系统的功能与特点的取样)之后, 由一部分评价者进行启发式评价, 另一部分评价者选取少量被试(3~5 人)执行简化的大声思维(大声思维是一种相对更有效的可用性测试方法[15]), 最后合并两部分评价的结果, 修改原初始设计。经过比较实验, 证明该方法比精细的可用性测试方法有更高的单位费用产出比。[16]

通过数学模型对用户的学习时间、使用绩效进行定量预测。用 GOMS 模型估计操作绩效、用用户已有知识的迁移等(TAMPE 模型)推算用户学习界面的时间;[11] 有人(Huguenard, Brian R)甚至已尝试通过菜单结构、用户短时记忆及任务特性定量预测用户使用该界面的绩效(运用 PBI 计算模型)。[17] 这些努力可能会弥补 IM 较难定量推断用户绩效的缺陷, 进而在界面设计阶段就估计用户的学习时间与操作绩效, 免去了为比较不同设计优劣必须分别进行可用性测试的麻烦。

### 2.3 方法的运用效率不高及其解决思路

从以上的 IM 看, 大都依赖评价人员的直接

手工评价;为解决这一问题,计算机支持的可用性评价方兴未艾,如 UIMS(用户界面管理系统)、ITS、CAUSE、及 MUSIC(包括 DRUM)等先进的用户界面设计与管理软件包中均具备了支持可用性评价的功能(如能自动判断菜单树是否层次太深、记录用户出错信息等)。[11, 12]这些正在发展的能与系统界面设计同步的智能计算机辅助评价系统(类似 CASE 工具),一则可实时自动指出哪一部分设计不够合理,二则可以为及时改进界面提供帮助选项,大幅度提高评价的速度与效率,降低对评价人员的评价经验要求。这也是今后可用性评价方法新的发展方向。

## 3　总结

本文介评、比较了最近十年来七种诊查型用户界面可用性评价方法,并提出了方法间进行定量比较的新思路。最后,对现有诊查型方法存在的问题提出了建议与趋势分析。

随着我国计算机科学的发展,我们希望有更多的科技人员能够投入到这一领域中,为实现用户界面设计、评价的本土化而努力。

## 参 考 文 献


[1] Neilsen J. Usability Engineering[M]. Boston: Academic Press, 1993. 16-21, 71-267.

[2] Robert AV. Usability Inspection Methods in Handbook of Human Computer Interaction(2nd)[M]. Helander M, et al. (eds), Elsevier Science B. V. 1997: 705-714.

[3] Alan JD. Human Computer Interaction[M]. Prentice Hall Europe: 1998. 405\|443.

[4] Barry Kirwan. Human error identification in human reliability assessment. Part2: Detailed comparison of techniques[J]. Applied Ergonomics, 1992, 23(6): 371-381.

[5] 王重鸣. 心理学研究方法[M]. 北京: 人民教育出版社, 1990. 90-98.

[6] Baber C, Stanton NA. Human error identification techniques applied to public technology: predictions compared with observed use[J]. Applied Ergonmics, 1996, 27(2): 119-131.

[7] 杨治良, 乐竟宏. 实验心理学[M]. 上海: 华东师范大学出版社, 1989. 105-115.

[8] 郑人杰. 实用软件工程[M]. 北京: 清华大学出版社, 1996. 38, 144.

[9] Berg. Gordon Lester Interface Design Guidelines For World Wide Web Planning Initiatives[D]. University Of Calgary (Canada) (0026) Essay for Medes Degree, 1997.

[10] Barry Kirwan. Human error identification in human reliability assessment——part1: detailed comparison of techniques[J]. Applied Ergonomics, 1992, 23(5): 299-318.

[11] Mohamed Khalifa. Computer-assisted evaluation of interface designs[J]. The DATA BASE for Advances in Information Systerms- Winter, 1998, 29(1): 66-81.

[12] Ben Shneierdman. Designing the User Interface: Strategies for Effective Human Computer Interaction (2nd)[M]. Addison-Wesley Publishing Company, 1995. 362-381.

[13] Karat J. User-centered Software Evaluation Methods in Handbook of Human Computer Interaction (2nd)[M]. Helander M, et al. (eds), Elsevier Science B. V, 1997. 689-704.

[14] Neilsen J, Mark Heuristic. Evaluation in Usability Inspection Methods[M]. Neilsen J (eds), NY Wiley and Sons. 1994: 1-7, 36-61.

[15] Henderson RD, Smith MC, Podd J, et al. A comparision of the four prominent user-based methods for evaluating the usability of computer software[J]. Ergonomics, 1995, 38(10): 2030-2044.

[16] Neilsen J. Guerrilla HCI: Using Discount Usability Engineering to Penetrate the Intimidation Barrier [M]. In Cost-justifying Usability Randolph G. Bias (eds), Academic Press, 1994, 245-272.

[17] Huguenard. Working Memory Failure In Human-Computer Interaction Modelin And Testing Simultaneous Demands For Information Storage And Processing (User Errors, Interface Design)[D]. Brian R. Carnegie-Mellon University for Degree: Phd 1993 Pp: 204 Source, 1994.

[18] Brooks J. Adding Value to Usability Testing in Usability Inspection Methods[M]. Neilsen J (eds), N.Y Wiley and Son, 1994. 258-265.

[19] Neville Stanton, Mark Young. Ergonomics Methods in Consumer Product Design and Evaluation, in Hurman Factors in Consumer Products[M]. Tayor & Francis Ltd, 1998. 21-52.

[20] 吴昌旭, 张侃. 诊查型用户界面可用性评价方法(IM)(上)——简介与评价[J]. 人类工效学, 2000, 6(3): 54-57.


[收稿日期] 2000-07-27