

可用性测试中的出声思维法

周荣刚^{1,2},张侃¹

(1. 中国科学院心理研究所,北京 100101;

2. 中国科学院研究生院,北京 100039)

摘要:出声思维法是一种很重要的可用性测试方法,得到了广泛的应用。但由于缺少统一的理论指导,不同使用者在使用的步骤、给予被试的指导语和干扰被试操作的程度等方面也存在诸多差异,因此很难实质性地评价出声思维法在可用性测试中的有效程度。文献回顾表明言语交流类型理论结合了 Ericsson-Simon 模型和可用性测试的具体特点,相比之下更能有效地为出声思维的实施提供指导。在此基础上对该方法进行了总结。

关键词:可用性测试;出声思维法;言语交流的类型理论

中图分类号:B849:TB18 **文献标识码:**A

1 概述

Ericsson 和 Simon 在认知心理学中提出了口语报告法,旨在研究人在问题解决中的认知过程,但其所基于的基本假设和形式为口语报告法在其他相关领域中的应用提供了基础和可能。出声思维(thinking aloud)在可用性测试(usability testing)中的应用便是一例,并已成为最常用的可用性测试方法^[1],甚至被认为是“唯一的最有价值的可用性工程方法”^[2]。该方法的基本特点是要求被试在完成任务的同时以口头言语的形式报告出任务操作情况,可用性测试专家通过对被试报告内容的分析,可以获知被测系统中存在的问题、哪些部分常为被试所忽视或误解以及被试对使用系统的看法等。本文从理论基础、使用中应注意的问题和相关的测试方法对出声思维法进行了总结,并简单回顾了相关的研究工作。

2 出声思维法应用于可用性测试中的理论基础

2.1 Ericsson 和 Simon 的口语报告模型

出声思维早期的理论基础源于 Ericsson 和 Simon 对口语报告法的诠释。口语报告法在人的信息加工模型提出之后获得了突破性的发展^[3]。Ericsson 等人根据任务操作过程中所介入的与任务无关的加工特点,对口语报告进行了三个层次水平的定义^[4]:水平 1,报告前无需进行转化的口

语报告,即报告的内容与短时记忆(STM)中编码的形式一致,思维序列没有发生,所以报告的内容最为可靠;水平 2,需要转化才可对思维过程进行描述的口语报告,转化的过程即为从 STM 中提取相关信息进而口语化的过程,这一过程会引起任务完成速度的下降,但是没有新信息的介入,思维序列和结构也没有发生变化,所以报告的内容也是可靠的;水平 3,需要额外加工介入才可进行的口语报告,额外的加工过程包括过滤性的加工(如只报告跟某主题有关的想法)、被试关于自身认知的推理以及对外界影响(如主试的提醒等)的感知,这些内容的报告要同时从 STM 和 LTM(长时记忆)中提取信息。Ericsson 和 Simon 认为对这部分内容的分析反映不出人操作时的认知加工过程,反对使用该水平报告的数据。

2.2 传统口语报告法在可用性测试中的问题

可用性测试中的出声思维有别于认知心理学研究中的口语报告法,主要体现在^[1,5]:出声思维旨在发现系统设计中存在的问题,使系统设计更符合人的认知加工过程,测试者关心的问题比较具体,建立人的认知加工模型不是可用性测试的重点;对被试与情境的交互状况进行分析也有助于认知过程的揭示,可用性测试关注的就是人机交互过程中系统存在的问题,这为解释出声

基金项目:中国科学院研究生科学与社会实践资助专项创新研究类项目(200314);教育部“十五”规划课题(FB011067);国家自然科学基金资助项目(30270465);中国科学院院长基金资助项目(JHJ02013)

作者简介:周荣刚(1978-),男,安徽当涂人,博士研究生,主要从事工程心理学方面的研究工作。

思维的数据提供了更多的交互情境；可用测试中被测界面很多时候只是半成品、甚至是设计早期的模型，出声思维随时会因此而中断，表现出更多的变异性。基于这些特点，Boren 等人^[1,5]通过对测试人员使用出声思维的现场观察，发现测试者并没有按照传统口语报告模型在应用出声思维法，主要体现在：测试者对被试的解释及其对被评测系统的喜好评价甚至修改建议更感兴趣，而这是传统口语报告中水平3的内容；由于被测系统往往是半成品，提醒被试不断“出声思维”不再是测试过程中主试与被试唯一的交互。尽管可用性测试者不断引用 Ericsson 和 Simon 的模型来说明其所收集数据的有效性，但实际应用已与传统口语模型大相径庭。不同的测试者在使用的步骤、给予被试的指导语和干扰被试操作的程度等方面也存在诸多差异，因此就很难实质性地评价出声思维法在可用性测试中的有效性。出声思维在可用性测试中的应用亟待新理论的指导。

2.3 言语交流类型理论

根据言语交流理论，言语交流中言语表达者（可用性测试中的被试）和倾听者（可用性测试中的测试者）虽然处于不对称的角色，但同起积极的作用。任何形式的言语活动都是有所指向的，表达者期待倾听者的反应，不可能像 Ericsson 和 Simon 所认为的那样“忽视”倾听者的存在而“自言自语”。Boren 等人认为言语交流类型理论更适宜为可用性测试中的出声思维提供理论指导：为避免测试者的干扰，在可用性测试中有必要建立高度非对称性的言语表达者——言语倾听者关系（即受测试者言语表达几率最大化而测试者言语表达几率最小化），而这种关系及其所依存的情境有益于言语类型（speech genres）的确定^[1,5]。言语类型是指支配特定范围内言语交流的一系列规则，包括言语交流中被试的角色、发生的环境和时间等。换言之，个体所采取的言语类型取决于交流的参与者、内容、地点、时间和原由，这些因素综合起来决定了交流双方所期望的最适宜反应。言语交流类型理论充分肯定了言语表达者和倾听者间的互动关系，而且测试者可根据具体情况设定不同的言语类型，进而使测试情境更为自然。

言语交流理论和传统口语报告法的目标是一致的，并结合了 Ericsson-Simon 模型和可用性测试的具体特点，相比之下更能有效地为出声思维的实施提供指导。

3 出声思维使用中应注意的问题

3.1 角色定义

根据言语交流理论，测试前对被测对象（软件界面、产品等）、被试和测试者的角色要作明确界定^[1,5]；被测系统是测试的对象，目的在于发现被测系统是否适合人的使用，这点要向被试强调；被试应被视为专家、是主要的言语表达者，这有利于被试给予测试任务更多的注意力、而减少对测试情境新颖性及其对潜在压力的关注；虽然测试者作为言语类型的创设者决定了测试的情境，但测试过程中测试者应作为学习者和倾听者而存在。所有角色的定义、包括测试环境和时间的安排都需要在测试前确定并在测试过程中贯彻下来。

3.2 如何使被试更好地进行口语报告

无论是传统口语报告法还是言语交流指导下的出声思维都强调要尽可能地保证数据的自然性、无干扰性和恒定性^[1,2]。测试者应尽可能少地干扰被试的操作，但诸如“嗯、哦、啊”等应答词的使用可给予表达者某种程度上的反应。应答词的选择及其使用的频率受很多因素的影响，很难确定，但要力求使整个测试过程更为自然。有研究表明英文中赋有疑问意味的“mm hm”和“uh-huh”最适合用于与被试的交流^[1,2]。

当被试忘记对操作进行报告时，提醒、鼓励被试继续报告时一般应遵循以下两点^[6,7]：集中于任务本身而非被评测系统的特征，如“你下一步的目标是什么？”而非“你觉得正在浏览的界面如何？”；集中于问题而非答案，如“你刚才说这个菜单有些混乱，是因为……？”而非“你觉得菜单混乱是不是因为×××造成的？”。总之，不可揣测被试意图或介入测试者个人的主观看法^[2]。

3.3 如何处理测试中的“突发事件”

可用性测试中意外事故或“非预期事件”较多，采用处理的方式应具有针对性^[1]。当出现由于被评价系统崩溃、程序中严重的缺陷或原型（Prototype）不完整而导致的“意外事故”时，要向被试强调故障是由被测系统本身所引起的，而不是由于其操作不当所致，故障处理后尽快确定继续测试的开始阶段，并尽可能快地使测试者对测试进程的控制降到最小。被试“提前”或以更为有效的方式完成任务，但忽视了被测系统所支撑的重要功能，此时可以用诸如“只有当……，这个任务才算完成”、“以这种方法完成这个任务比较合理且有新意，你还能用其他方法实现这个任务

吗?”等提示被试。“非预期事件”还包括被试询问测试任务中已有答案的问题、询问的问题正是需要他们完成的任务以及被试存在较为严重的进行与测试无关的“谈话”倾向。处理时要较好地保持被试为“专家”和测试者为“学习者”的角色,尽可能把原因归于被测系统。

4 研究进展及相关的出声思维法

出声思维法具有所需被试人数少^[8]、费用较低等优点,适合系统的设计、测试和发布阶段,目前有关的研究主要体现在:探讨出声思维所基于的理论模型^[1],研究者试图使被试以更自然的状态进行出声思维,言语交流类型理论作为对 Ericsson-Simon 模型的补充便是一很好的尝试。这一趋势与认知领域研究者不断从理论上(如逐渐重视水平 3 的数据收集^[9]、从情境的观点重新诠释“出声思维”^[10])对口语报告法进行改进的努力是一致的。出声思维是否会改变或如何改变操作绩效方面的研究^[2,11],被试口语报告前需要进行思维组织,完成任务的速度和正确性会随之有所改变,但改变的程度和方向目前没有取得一致的看法。任务设计方面的研究:任务设计是出声思维过程中一个重要的基本程序,不同的任务设计会影响到出声思维的数据。在用出声思维评价网站可用性时,不同任务设计(如简单搜索任务、申请任务和预测任务)会影响对同一网站的可用性评价结果,任务的设计应与所提出的评价目标相一致^[12,13]。

研究进展还表现为测试人员在方法上所进行的尝试,Neison 总结了三种跟出声思维法相关的测试方法。建构性交互(constructive interaction)^[2]也叫共同发现学习(codiscovery learning),即两个被试一同使用被测系统,通过分析二人的言语交流发现系统存在的问题。这种更适宜被试为儿童(以标准出声思维法很难使儿童完成测试任务)的测试。测试情境更为自然,但两个被试解决问题的方法可能完全不同,甚至不能共同完成测试任务。回顾性测试(Retrospective Testing)^[2,11]:即先以某一方法检测被试任务操作的情况,任务完成后再让被试边看录像边进行追述性口语报告。回顾性测试中被试倾向于对操作加以解释和说明,没有任务情景下的压迫感而测试者还可以就某一细节进行询问,可以获得更多信息,在难以找到目标被试时更有价值,但其所用时间几乎是标准大声思考的两倍,费用相对较高。

教练法(Coaching Method)^[21]:测试中教练竭力引导被试正确使用被测系统完成测试任务,进而收集数据的方法。测试者本身和专家用户都可以充当教练,被试可以向教练询问任何跟系统有关的问题,教练要尽其所能回答。教练法针对的是新手用户,旨在发现隐藏在其所提问题背后的关键信息,从而在训练、文档和界面中提供这样的信息。当由于系统/界面的用户群小、专业性强和高收入而导致的被试难以寻找时,教练法更为适用。

参考文献:

- [1] Boren MT, Ramey J. Thinking aloud: reconciling theory and practice[J]. IEEE Transactions on Professional Communication, 2000, 43(3): 261-278.
- [2] Nielsen J. Usability engineering[M]. Cambridge, MA: AP Professional, 1993.
- [3] 王重鸣. 心理学研究方法[M]. 北京:人民教育出版社, 1990.
- [4] Ericsson KA, Simon HA. Protocol analysis: verbal reports as data[M]. Cambridge, MA: MIT Press, 1993.
- [5] Ramey J, Boren MT. Keep them thinking aloud two ways to conduct a verbal protocol [EB/OL]. <http://www.up-assoc.org/conf2001/downloads/aloud.pdf>, 2003-05-01.
- [6] Barnum CM. Usability testing and research [M]. New York: Longman, 2002.
- [7] Tamler HM. How (much) to intervene in a usability testing session[J]. Common Ground, 1998, 8(3): 11-15.
- [8] Nielsen J. Estimating the number of subjects needed for a thinking aloud test [J]. International Journal of Human-Computer Studies, 1994, 41(3): 385-397.
- [9] 任洁, 许尚侠. 当代心理学对口语报告的研究评述 [J]. 心理科学, 1998, 21(1): 77-78.
- [10] Shu Ching Yang. Reconceptualizing think-aloud methodology: refining the encoding and categorizing techniques via contextualized perspectives[J]. Computers in Human Behavior, 2003, 19(1): 95-115.
- [11] Dumas JS, Redish JC. A practical guide to usability testing [M]. Exeter: Intellect Books, 1999.
- [12] Benbunan FR. Using protocol analysis to evaluate the usability of a commercial web site [J]. Information & management, 2001, 39(2): 151-163.
- [13] Waes LV. Testing usability tests for web sites: an evaluation of task variation in think aloud sessions [A]. Michitake H. Human computer interaction-interact '01 [C]. Amsterdam: IOS Press, 2001. 771-772.

[收稿日期]2004-07-14

[修回日期]2005-01-13