

人机交互过程中认知负荷的综合测评方法^{*}

李金波 许百华

(浙江大学心理与行为科学学院, 杭州 310028)

摘 要 设计模拟网络引擎搜索和心算双任务实验,分析主观评定、绩效测量和生理测量三类评估指标对认知负荷变化的敏感性;采用因素分析、BP神经网络和自组织神经网络三种建模方法,探索人机交互过程中认知负荷的综合评估建模方法。结果显示:心理努力、任务主观难度、注视时间、注视次数、主任务反应时、主任务正确率6个指标对认知负荷变化敏感;采用多维综合评估模型对双任务作业认知负荷进行测量总体上比采用单一评估指标的测量更为有效。BP网络和自组织神经网络两种神经网络模型对认知负荷的测量结果优于传统的因素分析方法。

关键词 人机交互; 评估; 认知负荷; 神经网络; 建模

分类号 B841.2

1 前言

随着计算机和自动化技术的发展,人在人机系统中的职能发生了很大变化,从原先执行简单的手工操作转变为主要从事监测和调节、控制系统运行状态的监控任务,人的认知负荷(Cognitive Load)大大加重。尤其是在军事、航空航天、核能等领域的复杂人机系统中,作业人员往往需要在短时间内处理大量的信息并快速作出反应与决策,容易出现认知负荷过高,甚至超负荷现象。过高的认知负荷将严重影响人的工作效率、操作可靠性和身心健康,进而影响整个人机系统的效率和可靠性。因此,认知负荷已成为人机系统评价的主要指标之一。

最早对认知负荷开展研究的是美国心理学家Miller(1956)。此后,国外一些学者对认知负荷的含义、结构和测评方法进行了大量的研究。Cooper(1990)将认知负荷定义为在特定的作业时间内施加于个体的工作记忆的心理活动总量。Paas和van Merriënboer(1994)认为认知负荷由多维度构成,是执行一项具体任务时施加于个体认知系统的负荷。认知负荷目前尚不能直接进行测量,而只能用间接的方法对其进行评估。常用的评估技术有任务绩效测量、主观评定和生理测量等三大类(Paas & Merriënboer, 1994; Brunken, Plass & Leutner, 2003;

Fairclough, Venables & Tattersall, 2005; Chris, Daniel & Alan, 2007; Ahlstrom & Friedman-berg, 2006)。其中任务绩效测量又分为主任务测量和次任务测量两种。主任务测量是通过直接测定作业者完成指定作业的绩效来评价该作业的认知负荷要求。它的基本假设是,随着作业难度的提高,对人的信息加工资源的需求相应地增加,当认知负荷对个体的要求超出其可支配的资源总量时,作业绩效将出现不同程度的下降。主任务测量常用的指标有操作准确性、反应时、信号漏失率和虚报率等。次任务测量采用双任务作业情境,要求作业者除执行主作业外再完成一项额外的作业(次作业),通过考察双任务作业情境下次作业绩效受影响的程度(相对于单任务情境)来间接评价主作业的认知负荷。生理测量是通过测定作业者在进行指定作业过程中出现的生理反应来间接地评估认知负荷,常用的生理指标有瞳孔直径、心率和事件相关电位等(Chris, Daniel & Alan, 2007; Ahlstrom & Friedman-berg, 2006)。主观评定技术是由作业者根据主观感受与体验来评估认知负荷,通常涉及个体知觉到的心理努力、任务难度和时间压力等方面。上世纪80年代以来,国外一些研究者相继开发出了多种认知负荷主观评定方法,如Paas和Van Merriënboer(1994)使用九级评定量表测量被试在理解任务材料上投入心理努力的程

收稿日期:2007-11-13

^{*}全国教育科学“十一五”规划教育部重点课题(DCA080141)、浙江省自然科学基金资助项目(Y604275)、浙江省哲学社会科学规划后期资助项目(08HQZZ038)资助。

通讯作者:许百华, E-mail: bhu305@zju.edu.cn

度,进而对认知负荷进行评估。Kalyuga, Chandler 和 Sweller(1997)使用由作业者自我报告任务难度的方法来测量认知负荷。此外,Salomon(1983)、Tabbers(2004)、Paul(2006)等也分别提出了相类似的认知负荷主观评定方法。

已有的研究表明,上述3类测评技术各有优点和局限性。例如,主任务测量比较直接,对作业本身无干扰,但只在中等负荷水平时较敏感;次任务测量相对比较敏感,效度较高,但容易对主任务产生干扰。此外,任务绩效与认知负荷的关系不一定是简单的线性关系。例如,Meister(1976)将认知负荷等同于任务需求,发现任务需求与工作绩效之间是曲线的关系;而 Waard(1996)则认为,任务需求与任务绩效之间呈“倒U”形的关系。主观评定无干扰,操作简便,但要求被试采用内省的方法来评定负荷水平,因而容易出现较大的偏差。此外,主观评定结果与个体特征、认知策略以及个体的心理和生理状况等多种因素有关,即使同一个人对同样的任务,在不同的作业时间段也有可能出现不同的评定结果。以 Paas 和 Van Merriënboer(1994)提出的心理努力问卷为例,该问卷看起来能够测量投入努力的主观知觉,但是却无法知道这种心理努力与真实认知负荷的关联程度。如有研究表明,低努力投入可能是由于任务过难,个体减少心理努力的结果(Reed, Burton & Kelly, 1985)。同样, Kalyuga, Chandler 和 Sweller 使用让个体自我报告任务难度差异方法所得出的结果也很可能是由于所需要完成的任务难度、个体个人能力水平等原因造成的。生理测量指标具有客观性和实时性,但是它们与认知负荷之间只是间接的联系,并且也可能会受到注意或情感等方面因素的影响(Brunken, Steinbacher & Plass, 2002)。因此,由认知负荷引起的某一生理指标的变化可能会被其它因素放大或缩小;而且不同的任务可能会产生不同的生理反应,一项生理指标对某一类任务适用,而对另一类任务则可能不适用。一般认为,不同测评技术分别适用于不同的情境、不同的负荷水平范围。因此,利用多种技术对认知负荷作综合评估以替代基于单一方法或指标的评估是比较合理的选择;同时,认知负荷的多维度特性也决定了对其作综合评估的必要性。近年来,国外一些学者运用多指标综合评估方法在心理工作负荷(Mental Workload)的测评中进行了一些探索性研究。这些研究以多个单一评估指标的测量结果作为依据,采用一定的建模技术对心理工作负荷进行多维度综

合评估。目前在这一领域较常用的建模技术主要有因素分析、回归分析和人工神经网络建模等。例如, Kilseop 和 Myung(2005)运用主成分分析将三种生理指标(脑电、眼动和心率)和主观负荷组合成一个综合评估指数,发现综合指数较单项指标能更准确地区分不同难度任务中被试的负荷水平。此外,他们还通过回归分析来考察主观负荷与生理指标之间的关系,以生理指标的测量结果来预测被试的主观评定的负荷状况。人工神经网络简称神经网络,在国外有关的研究中,运用神经网络模型对负荷进行多维度综合评估主要有两种模式。一种模式是将主观负荷、作业绩效指标或生理指标的测量结果作为网络的输入项,通过建立不同的神经网络模型对负荷进行评估或等级划分。例如, Roger, Rebecca 和 Gory(2003)在一项模拟多任务作业的实验中,收集了5个绩效指标作为输入项,建立起BP神经网络模型。经检验,该模型对预测不同任务条件下被试的负荷状况具有较高的准确率(在66.9%~96.0%之间)。Mazaeva 和 Ntue(2001)在一项飞行作业的模拟实验中对被试进行了30个EEG指标的测量,然后运用主成分分析法将这些指标归并为少数的几个指标,以它们作为网络的输入项建立起自组织神经网络模型。结果显示,借助自组织神经网络模型能比较准确地预测信息加工要求不同的任务中负荷发生的变化。此外,其他的研究者利用相类似的方法也得到了较理想的结果(Laine, Bauer & Jr Tefrey, 2002; Noel, Bauer & Lanning, 2005; Grootjen, Neerincx & Veltman, 2006; Shayeghi, Shayanfar, 2006; Shayne, Penelope & Andrew, 2007)。第二种模式是将影响负荷的主要因素作为输入项,通过建立神经网络模型,对不同任务难度条件下被试的负荷状况进行评估。例如, Ding 和 Sheue(1999)将影响监控作业的跨度(Span)、差异度(discriminate)、可预测度(predict)和注意转换(transfer attention)4个因素作为输入项建立BP神经网络对负荷进行评估,并将神经网络评估结果与负荷上限值进行比较,超过上限值就认为是超负荷,并据此对监控作业过程中的负荷状况进行实时分析和任务的动态分配。但从总体上来看,国外学者利用神经网络对负荷进行评估或等级划分还处于起步阶段,并且已有的研究中,任务条件各异,神经网络输入变量不同,因而所得的结果也难以进行比较。此外,有的研究者还运用多层次模糊综合评价方法对负荷进行综合评估研究(Currie, 1997)。

在认知负荷综合评估方面,国外学者也进行了一些尝试,如 Paas 和 van Merriënboer (1993)、Tuovinen 和 Paas (2004) 将心理努力与主任务绩效这两类指标结合起来对认知负荷进行综合测评,获得了利用单一评估指标无法得到的许多重要信息。他们还发现,采用多种技术对认知负荷进行多维综合评估可显著提高评估敏感性。但到目前为止,对认知负荷的综合评估研究还是零碎的,缺乏建模方法的系统研究和不同方法之间的比较研究。在国内,目前还未见这方面的研究报道。

借鉴国外已有的多指标综合评估方法在心理工作负荷测评中的运用,本研究拟结合复杂人机系统中视觉信息加工作业的特点,通过模拟实验,分析主观评定、绩效测量和生理测量三类评估指标对认知负荷变化的敏感性,然后基于经筛选的认知负荷评估指标的测量结果,分别采用因素分析、BP 神经网络和自组织神经网络三种建模方法,构建双任务作业中认知负荷的多维综合评估模型,并对这三种综合评估模型进行比较,从中选出最合适的建模方法。

2 方法

2.1 实验设计

采用双任务作业实验,主任务为模拟的网络引擎搜索(Chen, 2006)。每次实验时,在计算机屏幕上相继呈现二幅画面,第一幅画面呈现搜索目标,包括要搜索的产品名称和两个指定特征(如形状、颜色);第二幅画面呈现搜索结果,包括 5 个与搜索目标同类型的产品,每个产品包含 4 个方面的特征(如品牌、颜色、形状和价格)。告知被试 5 个产品中有 1 至 4 个符合搜索目标要求的目标项,要求被试查找出所有的目标项,同时记住目标项的另外二个特征(如品牌和价格);第二幅画面呈现时间为 16 秒,被试完成搜索后按键或在 16 秒后第二幅画面消失,被试口头报告搜索到的各个目标项的另外二个特征。次任务作业为心算作业,即在呈现第二幅画面后随机呈现一个心算问题,要求被试作两个个位数相加而后除以 3 的心算,并对两数之和能否被 3 整除在计算机键盘上尽快作出相应的按键反应。

主任务分成 3 种复杂水平,通过改变第二幅画面 5 个搜索结果中包含的目标项数目来加以控制,分别为 5 个搜索结果中包含 1、2 和 4 个目标项,代表低、中、高三种主任务复杂水平。实验采用被试内设计,每种实验任务条件下实施 8 次搜索和心算作业,持续时间约 3 ~ 5min,不同实验任务之间休息 3

分钟。

2.2 被试

共 33 名大学生和研究生参加了实验,其中男生 15 名(占 45.5%)、女生 18 名(占 54.5%);被试年龄在 20 至 27 岁之间,平均年龄为 22.4 岁,所有被试视力正常,优势手均为右手。

2.3 实验仪器和测试工具

实验在计算机上进行,眼动测量采用美国应用科学实验室(Applied Science Laboratory, ASL)生产的 ASL 504 非头盔式眼部跟踪记录仪。

实验中认知负荷主观评定分心理努力和任务主观难度二个方面。心理努力评定采用 Paas 等人(1994)制定的量表,经 Paas 验证,该量表的信度系数为 0.90(Paas & Van Merriënboer, 1994)。由于“心理努力”是一个专业术语,对被试来说,可能不容易理解;为此,我们在正式实验的量表中采用“完成任务的努力程度”来表示心理努力。量表为 7 级评分制,要求被试根据自己的感受从 1 ~ 7 中选择一个合适的数字,从 1 到 7 努力程度依次递增,1 表示非常轻松,4 表示中度努力,7 表示非常努力。任务主观难度采用 Kalyuga, Chandler 和 Sweller(1999)制定的量表,同样采用 7 级评分制,从 1 到 7 任务难度依次递增,1 表示非常容易,4 表示中等难度,7 表示非常困难。研究表明,使用该量表能够较准确地测查出认知负荷的变化。

本研究利用 E-prime 实验心理专用软件编制程序,实现主任务、次任务呈现时间的控制和测量相应的反应时等数据。

2.4 实验程序

实验对每个被试单独进行。实验前,被试完成 8 次单任务的心算作业,以这 8 次心算的平均时间作为被试的心算作业反应时的基础测量值(BM)。接着,安排被试仔细阅读计算机上呈现的指导语,然后进行两次双任务作业的练习,主任务为网络搜索作业,次任务为心算作业。

正式实验中,主任务分低、中、高 3 种任务复杂水平。在每种任务复杂水平的条件下。被试按指导语进行 8 次由网络搜索与心算相结合的双任务作业。实验完成后,被试在心理努力和任务主观难度两个量表上对认知负荷大小作出主观评定。主任务的 3 种复杂水平的实验顺序在被试之间做了平衡。

2.5 数据收集与整理

每一轮实验完成后,收集和整理以下一些数据:主任务正确率:在每次试验后,被试准确口头报

告一个特征就计为一个正确回答,每种实验任务条件下 8 次试验正确回答的特征数占有所有正确特征数的比例就作为该实验任务条件下的主任务正确率。

主任务反应时:利用 E-prime 编制的程序自动记录每次试验中从搜索结果页面呈现到被试按键的时间,计算被试每种实验任务条件下完成 8 次试验的平均时间,以它作为该实验任务条件下的主任务反应时。

次任务正确反应时变化:在双任务条件下测得的心算反应时间称为被试双任务心算反应时测量值(DTM)。将 DTM 减去单任务心算反应时测量值,即 $CWR = DTM - BM$,计算 8 次试验中回答正确的 CWR 平均值就是次任务正确反应时变化。

注视时间、注视次数、眼跳距离和瞳孔直径:利用 ASL504 眼动仪记录眼动指标数据。眼睛如果在较短的时间(100ms)稳定在相对的空间阈限范围内(1°),速度低于阈限(每秒 15 ~ 100°)记为一个注视

点。注视点的数量就是注视次数;每个注视点的平均所占时间就是注视时间;从当前注视点位置到下一个注视点位置之间间隔长度就是眼跳距离;每次注视时瞳孔的平均大小就是瞳孔直径。

心理努力和任务主观难度:每种实验任务条件结束后用主观评定量表分别获得心理努力和任务主观难度的评估值。

收集到的数据利用 SPSS 软件进行分类汇总,并进行探索分析,去除其中的奇异值。

3 结果与分析

3.1 任务复杂性对认知负荷的影响

利用 SPSS 软件计算各个认知负荷评估指标在不同任务复杂性条件下的均值,并分别对各个评估指标在不同任务复杂性条件下均值的差异进行方差分析,结果见表 1。

表 1 各评估指标在三种任务复杂性条件下的均值及变化的方差分析

评估指标	任务复杂性			<i>F</i>	<i>p</i>
	低 <i>M</i>	中 <i>M</i>	高 <i>M</i>		
心理努力	2.30	4.21	5.48	21.13	0.002
任务主观难度	2.33	3.73	5.15	20.14	0.003
注视时间	66.01	83.33	83.68	8.91	0.000
注视次数	283.82	366.91	388.55	18.73	0.000
眼跳距离	50.72	48.58	50.60	0.36	0.701
瞳孔直径	31.56	31.82	31.91	0.02	0.977
主任务反应时	12385.85	15444.44	15866.41	58.58	0.000
主任务正确率	0.92	0.77	0.55	117.44	0.000
次任务正确反应时变化	1607.15	1988.45	1404.87	1.07	0.346

从表 1 可以看出,随着任务复杂性增加,心理努力投入和任务主观难度增大,注视时间、注视次数、主任务反应时增加,主任务正确率降低,瞳孔直径变大,而眼跳距离和次任务正确反应时变化则没有表现出规律性的变化。方差分析结果显示,在不同的任务复杂性条件之间,心理努力、任务主观难度、注视时间、注视次数、主任务反应时、主任务正确率均表现出极显著的差异($p < 0.01$)。而眼跳距离、瞳孔直径和次任务正确反应时变化的差异不显著($p > 0.05$)。由此可见,对任务复杂性变化敏感的评估指标主要有心理努力、任务主观难度、注视时间、注视次数、主任务反应时、主任务正确率等。

3.2 基于 BP 网络的认知负荷综合评估模型

BP 网络是一种单向传播的多层前向网络,它能

在无需事先了解输入—输出模式映射关系数学方程的情况下,通过对样本进行学习训练,建立由 n 维输入项到 m 维输出项之间的非线性映射。BP 网络由输入层、输出层和隐含层构成,其具体设计包括输入/输出项设计、网络结构设计、网络训练与测试等几个方面。输出项神经元个数即所要区分和识别的认知负荷水平的等级数。在实际的作业情境中,认知负荷水平受任务特征、个体特征及两者交互作用的影响,但在某个特定的时刻或较短时间内,个体特征又具有相对的稳定性,而任务特征是决定认知负荷大小的关键因素(Paas & van Merriënboer, 1994)。参照国外同类研究,本研究通过改变任务特征(任务复杂性)来改变认知负荷水平,根据前述的实验设计,按照任务复杂性的不同将认知负荷分成 3 个

等级,与此相对应,BP 网络设计中输出项神经元个数设置为 3 个。输入项的确定实际上就是特征量的提取,对于特征量的选取,主要考虑它是否与认知负荷有比较确定的相关关系。由上述分析可知,与认知负荷显著相关的因素有心理努力、任务主观难度、注视时间、注视次数、主任务正确率和主任务反应时变化共 6 个。因此,输入项神经元个数设置为 6 个。网络结构中的网络层数采用单隐层的三层网络,根据初测结果,隐层神经元个数选择 14 个,隐含层传递函数采用 S 型的正切函数,输出层传递函数采用 S 型的对数函数,训练函数选择 trainlm 函数(Levenberg-Marquardt 算法),期望收敛精度为 0.001。随机选择 33 个被试中的 24 个共 72 组数据作为训练样本,其余的 9 个被试共 27 组数据作为测试样本。利用 MATLAB 软件编制程序进行网络训练和测试。训练之前,将所有输入项进行归一化处理,转化成 [0,1] 区间的数据,输出项用 0-1 二分数据形式表示。训练好的网络采用测试样本进行测试可得到测试负荷值,将测试负荷值重新转换成等级数据形式,就可以与测试样本中的输出项值(期望负荷值)进行比较。我们选择期望负荷值与测试负荷值之间的绝对误差、相对误差(取绝对值)和均方误差作为网络模型的性能指标,计算每个被试在每种负荷水平下测试负荷值与期望负荷值的绝对误差、相对误差绝对值和均方误差,求取每种负荷水平下所有被试测试负荷值与期望负荷之间绝对误差、相对误差绝对值和均方误差的均值;同时计算三种负荷条件下绝对误差、相对误差绝对值和均方误差的总均值,结果见表 2。

表 2 基于 BP 网络建模的综合评估结果

期望负荷值	测试负荷值	绝对误差	相对误差(%)	均方误差
1	0.83	-0.17	17.20	0.15
2	1.79	-0.21	10.40	0.22
3	2.93	-0.15	2.30	0.05
绝对误差均值 = -0.15				
相对误差均值(%) = 9.97				
均方误差均值 = 0.14				

注:表中期望负荷值 1、2、3 分别表示低、中、高三种认知负荷水平,以下同。

从表 2 可以看出,在不同负荷水平条件下,期望负荷值与测试负荷值的绝对误差、相对误差和均方误差不同。相对而言,BP 网络模型在高负荷水平时其综合评估结果最为准确(绝对误差、相对误差、均

方误差均最小)。

3.3 基于自组织神经网络(SOM)的认知负荷综合评估模型

自组织特征映射网络(SOM)是一类无教师、自组织、自学习网络,它能将输入项有序地映射在输出层上,实现将样本的高维空间在保持结构不变的条件下降投影到低维空间。因此,可以应用于认知负荷水平的等级划分和识别研究中。

SOM 网络由输入层和竞争层组成,我们同样选择与认知负荷有显著相关的 6 个单维度评估指标作为网络的输入项。输出层的设计涉及神经元个数和神经元排列的设计两个方面。神经元个数设定为 3 个,本研究仅对认知负荷的水平进行等级划分,因而神经元的排列采用一维线阵来表示。训练样本与测试样本与 BP 网络相同,经测试,训练次数达到 1000 次以上时,训练结果已差异不大,因而训练次数选择 1000 次,对训练好的网络进行测试,结果见下表 3。

表 3 基于 SOM 网络建模的综合评估结果

期望负荷值	测试负荷值	绝对误差	相对误差(%)	均方误差
1	1.10	0.10	10.00	0.10
2	2.10	0.10	25.00	0.50
3	2.80	-0.20	6.67	0.20
绝对误差均值 = 0.00				
相对误差均值(%) = 13.89				
均方误差均值 = 0.27				

从表 3 中可以看出,利用构建 SOM 网络模型对认知负荷水平进行分类其相对误差要高于 BP 网络模型。

3.4 基于因素分析的认知负荷综合评估模型

利用因素分析可以从众多的评估指标中抽取少数几个综合指标(公共因素),并根据特征根和因素负荷计算被试在公共因素上的得分。由于每个公共因素的方差贡献率表示该因素对总体变异的解释程度,反映了该因素的相对重要性,因而可以作为各个因素的加权系数。计算被试的因素得分的加权总和,这个总和值就可以作为认知负荷的综合评估指标值(王权,李金波,2002)。

选择上述 6 个认知负荷评估指标,利用主成分分析方法抽取公共因素,当抽取公共因素为 3 个时,累计方差贡献率已达 89.86%。经方差极大正交旋转后其因素负荷矩阵见下表 4。

表 4 评估指标的因素分析结果

评估指标	公共因素		
	1	2	3
V1 心理努力	-0.76	-0.36	0.15
V2 难度感知	0.94	0.02	0.15
V3 注视时间	0.95	0.08	0.16
V4 注视次数	0.09	0.22	0.95
V5 主任务反应时	0.12	0.96	0.07
V6 主任务正确率	0.17	0.83	0.47
特征根	3.04	1.56	0.79
方差贡献率(%)	40.09	29.90	19.87

根据特征根、因素负荷,计算被试在 3 个公共因素上的得分,以每个公共因素的方差贡献率作为加权系数计算 3 个公共因素得分的加权总和。并将其一一转化为[1,3]区间内相应的值,以便与期望负荷值进行比较,结果见表 5。

表 5 基于因素分析的综合评估结果

期望负荷值	测试负荷值	绝对误差	相对误差(%)	均方误差
1	0.82	-0.18	51.18	0.32
2	1.79	-0.21	18.63	0.23
3	2.42	-0.58	19.48	0.47

绝对误差均值 = -0.32
 相对误差均值(%) = 29.76
 均方误差均值 = 0.34

4 讨论

4.1 评估指标对认知负荷变化的敏感性

目前,国内外学者对认知负荷的研究大多采用实验室实验方法,即通过控制任务的难度(复杂性)来改变认知负荷水平,在此基础上分析各种评估指标值随任务难度(复杂性)改变的规律,进而判断各个指标用于评定负荷状况的意义。本研究通过改变任务复杂性来控制认知负荷水平,并测量不同任务复杂性水平下的主任务与次任务绩效、生理反应和主观评估指标。以此为基础,探讨认知负荷与各个评估指标之间的关系。结果显示,心理努力、任务主观难度、注视时间、注视次数、主任务反应时和主任务正确率共 6 个评估指标对认知负荷的变化有显著的敏感性。

目前,在对心理努力与认知负荷(任务复杂性)的关系的解释上存在两种相互矛盾的假设,Gimino(2000)认为两者之间是正向的线性关系,而 Salomon(1983)则认为两者是“倒 U 形”的曲线关系。本研究结果支持 Gimino 的假设,心理努力与认

知负荷(任务复杂性)之间存在正向的相关(系数为 0.72,见表 6)。当任务非常容易时,由于个体利用自动化知识就可以完成任务,因而其认知负荷低,投入的心理努力少,绩效水平能保持在最高水平;当任务难度增大时,个体需要动用更多的陈述性知识,心理努力投入增加,认知负荷也相应增大,而绩效水平则会出现下降。因而绩效与认知负荷(任务复杂性)的关系是一种负向的相关关系,表 6 显示认知负荷与主任务正确率的相关系数为 -0.82;而与任务主观难度的相关为 0.69。注视时间、注视次数与认知负荷(任务复杂性)的关系也是一种正向的相关关系(相关系数分别为 0.39 和 0.53)。这说明认知负荷(任务复杂性)对任务即时加工有着重要的影响,认知负荷(任务复杂性)越大,注视时间越长,注视次数越多,反映个体在对任务加工过程中需要接收、编码、储存的元素数目越多。本研究的这一结果支持 Sweller 的认知负荷理论。

本研究中,眼跳时间、瞳孔直径和次任务反应时在不同的认知负荷(任务复杂性)水平下无显著变化,这与 Paas(1994)等人报道的瞳孔反应对认知负荷变化具有高度敏感性的实验结果不一致。但如果对本研究得出的不同任务复杂性下瞳孔直径的均值作一分析,可以发现,随着认知负荷(任务复杂性)增大,瞳孔直径还是表现出增大的趋势。眼跳距离与认知负荷(任务复杂性)的关系存在两头高、中间低的特点,即任务复杂性低或高时眼跳距离较大,任务复杂性中等时眼跳距离较小;而次任务正确反应时变化与认知负荷(任务复杂性)的关系则相反,呈现出两头低、中间高的特点。

4.2 单个指标测评方法与多维综合评估方法的比较

抽取测试样本在 3 种实验条件下的各个单一评估指标测评值,以及基于 3 种多维综合评估模型得出的评估值,分别计算他们与相应实验条件下认知负荷水平的 Spearman 相关,结果见表 6。

从表 6 可见,各项测评结果与认知负荷水平均存在显著相关;相比之下,BP 网络评估值与认知负荷水平的相关性最高(相关系数为 0.93),其次为 SOM 网络,而后为主任务正确率和因素分析评估。由此可见,基于综合评估模型的评估结果与认知负荷水平的相关总体上要高于单项评估指标测评值与认知负荷水平的相关,这说明采用多维综合评估模型对认知负荷进行评估总体上比采用单一评估指标更为有效。而在采用单项指标的几种测评方法中,

主任务正确率与认知负荷水平的相关性最高。

表 6 单个指标测评值、综合评估结果值与认知负荷水平值的相关

	心理努力	任务主观难度	注视时间	注视次数	主任务反应时	主任务正确率	BP 网络评估值	SOM 网络评估值	因素分析评估值
认知负荷水平	0.72 *	0.69 *	0.39 *	0.53 *	0.65 *	-0.82 *	0.93 *	0.82 *	0.81 *

注: * $p < 0.001$

4.3 三种综合评估模型评估结果之间的比较

对三种综合评估模型的评估结果进行整理,结果见下表 7。

表 7 三种综合评估模型评估结果的比较

模型	绝对误差均值	相对误差均值(%)	均方误差均值
BP 神经网络	-0.15	9.97	0.14
SOM 神经网络	0.00	13.89	0.27
因素分析	-0.32	29.76	0.34

由表 7 可以发现,在相对误差均值上,BP 网络的相对误差最低(9.97%),其次是 SOM 网络(13.89%),两者均明显低于因素分析(29.76%)。在绝对误差均值上,SOM 网络最低,BP 网络次之。而在均方误差方面,BP 网络最低,SOM 网络次之。可见利用神经网络模型对认知负荷作综合评估总体上要优于传统的因素分析方法。神经网络通过使用非线性映射关系来贴近复杂的非线性系统,摆脱了传统的线性分析工具,而更适用于认知负荷的测评。

5 结论

本研究通过分析任务特征(任务复杂性)与认知负荷评估指标(作业绩效、主观评估、生理指标)的关系,发现心理努力、任务主观难度、注视时间、注视次数、主任务反应时、主任务正确率这 6 项指标对认知负荷的变化具有较高的敏感性。以这 6 项指标的测量结果为依据,分别采用因素分析法、BP 神经网络和 SOM 神经网络建模技术构建认知负荷多维综合评估模型,发现在双任务情境下,采用多维综合评估模型对认知负荷进行评估比采用单一评估指标更为有效。而在 3 种多维综合评估模型中,BP 网络和自组织神经网络(SOM)两种神经网络模型对认知负荷的测量结果优于采用因素分析法构建的模型。

本研究由视觉搜索与心算组成的双任务实验结果来分析认知负荷评估指标的敏感性,然后基于对搜索作业负荷变化敏感的评估指标,建立起认知负荷的多维综合评估模型。然而,作业性质不同,认知

负荷评估指标的敏感性会不一样。因此,本研究建立的认知负荷综合评估模型不完全适用于各类人机系统中认知负荷的测评问题。本研究的意义在于提供了如何构建认知负荷多维综合评估模型的方法,并对几种建模方法作了比较。在实际运用这些建模方法时,对于不同性质的作业,需要对本研究建立的模型作出适当的修改。

参 考 文 献

- Ahlstrom, U., Friedman-berg, F. (2006). Using eye movement activity as a correlate of cognitive workload. *International Journal of Industrial Ergonomics*, 36, 623-636.
- Brunken, R., Plass, J., Leutner, D. (2003). Direct measurement of cognitive load in multimedia learning. *Educational Psychologist*, 38, 53-61.
- Brunken, R., Steinbacher, S., Plass, J. (2002). Assessment of cognitive load in multimedia learning using dual-task methodology. *Experimental Psychology*, 49 (1), 109-297.
- Chen, C. T. (2006). *Cognitive processing during web search: the role of working memory load in selective attention and inhibitory control*. unpublished Doctorial Dissertation, Indiana University.
- Chris, B., Daniel, L., Alan, Y. (2007). EEG correlates of task engagement and mental workload in vigilance, learning, and memory tasks. *Aviation, Space, and Environmental Medicine*, 78 (5), 231-244.
- Cooper, G. (1990). Cognitive load theory as an aid for instructional design. *Australian Journal of Educational Technology*, 6 (1), 108-113.
- Currie, N. J. (1997). *The development of a composite index of mental workload via fuzzy set applications*. unpublished Doctorial Dissertation, University of Houston.
- Ding, Y. L., Sheue, L. H. (1999). Use of neural networks to achieve dynamic task allocation: a flexible manufacturing system example. *International Journal of Industrial Ergonomics*, 24 (2), 281-298.
- Fairclough, S. H., Venables, L., Tattersall, A. (2005). The influence of task demand and learning on the psychophysiological response. *International Journal of Psychology*, 56, 171-184.
- Gimino, A. E. (2000). *Factors that influence students' investment of mental effort in academic tasks: a validation and exploratory study*. unpublished Doctorial Dissertation, University of Southern California.

- Grootjen, M. , Neerinx, M. A. , Veltman, J. A. (2006). Cognitive task load in a naval ship control centre: from identification to prediction. *Ergonomics*, 49 (12), 1238 – 1264.
- Kalyuga, S. , Chandler, P. , Sweller, J. (1999). Managing split-attention and redundancy in multimedia instruction. *Applied Cognitive Psychology*, 13, 351 – 371.
- Kilseop, R. , Myung, R. (2005). Evaluation of mental workload with a combined measure based on physiological indices during a dual task of tracking and mental arithmetic. *International Journal of Industrial Ergonomics*, 35, 991 – 1009.
- Laine, T. I. , Bauer, K. W. , Jr Teffrey, W. (2002). Selection of input features across subjects for classifying crewmember workload using artificial neural networks. *IEEE Transactions on Systems, Man, and cybernetics-part A: Systems and Humans*, 32 (6), 691 – 704.
- Mazaeva, N. , Ntuen, C. (2001). Self-Organizing Map(SOM) model for mental workload classification. *IEEE Transactions on Systems, Man, and cybernetics-part A: Systems and Humans*, 31 (6), 1822 – 1825.
- Meister, D. (1976). *Behavioral foundations of system development*. New York: Wiley Press.
- Miller, G. A. (1956). The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological Review*, 63, 81 – 97.
- Noel, J. B. , Bauer, K. W. , Lanning, J. W. (2005). Improving pilot mental workload classification through feature exploitation and combination: a feasibility study. *Computer & Operations Research*, 32, 2713 – 2730.
- Paas, F. , van Merriënboer, J. (1993). The efficiency of instructional conditions: An approach to combine mental effort and performance measures. *Human Factors*, 35, 737 – 743.
- Paas, F. , van Merriënboer, J. (1994). Instructional control of cognitive load in the training of complex cognitive tasks. *Educational Psychology Review*, 6 (4), 351 – 371.
- Paul, A. (2006). Using subjective measures to detect variations of intrinsic cognitive load within problems. *Learning & Instruction*, 16 (5), 389 – 400.
- Reed, W. M. , Burton, J. K. , Kelly, P. (1985). The effects of writing ability and mode of discourse on cognitive capacity engagement. *Research in the Teaching of English*, 19, 283 – 297.
- Roger, W. S. , Rebecca, L. G. , Gory, B. R. (2003). Model of Mental Workload. *International Journal of Industrial Ergonomics*, 28 (2), 381 – 407.
- Salomon, G. (1983). The differential investment of mental effort in learning from different sources. *Educational Psychologist*, 18 (1), 42 – 50.
- Shayne, L. , Penelope, S. , Andrew, N. (2007). Modeling and Predicting Mental Workload in En Route Air Traffic Control: Critical Review and Broader Implications. *Human Factors*, 49 (3), 376 – 399.
- Shayeghi, H. , Shayanfar, H. A. (2006). Application of ANN technique based on μ -synthesis to load frequency control of interconnected power system. *International Journal of Electrical Power & Energy Systems*, 28 (7), 503 – 511.
- Tabbers, H. , Martens, R. L. , van Merriënboer, J. (2004). Multimedia in structions and cognitive load theorys :Effects of modality and chang. *Britian Journal of Educational Psychology*, 74, 71 – 81.
- Tuovinen, J. , Paas, F. (2004). Exploring multidimensional approaches to the efficiency of instructional conditions. *Instructional Science*, 32, 133 – 152.
- Waard, D. (1996). *The measurement of drivers' mental workload*. unpublished Doctorial Dissertation, University of Groningen.
- Wang, Q. , Li, J. B. (2002). *Confirmatory Factor Analysis*. Hangzhou China: Zhejiang University Press. (in Chinese)
- [王权, 李金波. (2002). 实证性因素分析. 杭州: 浙江大学出版社].

Synthetic Assessment of Cognitive Load in Human-Machine Interaction Process

LI Jin-Bo, XU Bai-Hua

(Department of Psychology and Behavioral Science, Zhejiang University, Hangzhou 310028, China)

Abstract

Cognitive load (CL) often becomes too great in modern human-machine systems, which affects the performance and reliability of system operators. Therefore an effective and reliable approach to measure CL is crucial. Currently there are three major measures of CL: these are subjective evaluation, performance measures and physiological assessment. None of these, however, can comprehensively reflect the state of CL under different task conditions. Thus, this study attempted to establish an appropriate multi-dimensional assessment technique to evaluate CL synthetically in the human-machine interaction process.

A dual-task experiment with simulated web-searching as the primary task and mental arithmetic as the secondary task was conducted. The primary task had three levels of complexity, high, medium, and low. Thirty-two subjects were selected to perform the dual-task with each complexity level of the primary task. The order of performance of the three complexity levels was randomized across subjects. Four types of indices were obtained, including primary-task measures, secondary-task measures, subjective evaluations and eye movement measures. The synthetic assessment models were constructed via factors analysis, BP artificial neural network and self-organizing feature map, respectively. The results showed that: (1) The six indices, which are mental effort, perceived task difficulty, duration of fixation, number of fixations, response time and the percentage of correct responses in the primary task, were sensitive to the change in cognitive load. (2) In dual-task situations, the multidimensional synthetic assessment of cognitive load generated more valid results than those based on the single assessment index. (3) According to the indices of relative, absolute and average errors, the two ANN models (BP neural network and self-organizing feature map) showed a more accurate measurement of cognitive load than the traditional technique of factors analysis.

Key words Human-machine interaction; cognitive load; assessment; neural network; modeling