

认知行为实验研究中最佳素材容量的选择与确定： 多元概化理论应用

罗照盛 郭小军

(江西师范大学心理学院, 南昌 330022)

摘要 心理学实验首先必须保证结果的有效性和可靠性。实验素材容量是影响实验结果的关键因素之一。本文的目的是通过运用多元概化理论分析几个典型认知行为实验的素材容量对实验结果精度的影响, 从而探讨最佳的实验素材容量的确定问题。研究结果发现, 在 IAT 测验上, 相容与不相容任务的素材容量为 50 时最佳, 可靠性指数为 0.92; 在图-词干扰范式中, 素材容量以 48 时为最佳, 可靠性指数为 0.95; 对线索提示范式, 50%有效提示中, 有效与无效提示最佳素材容量为 35, 可靠性指数是 0.97。研究表明, 多元概化理论可以很好地用于确定认知行为实验中素材的最佳容量。

关键词 多元概化理论; 实验信度; 素材容量

分类号 B841

1 前言

心理学实验首先必须保证结果的有效性和可靠性。实验信度指的是实验结果和结论的可靠性、稳定性及一致性程度。如果实验结果是不可靠的, 那么研究结论将毫无价值。影响实验结果的因素非常多, 而现在认知行为实验已经发展出了许多精巧的实验研究范式及其变式, 在控制实验误差方面也已经考虑得非常周到。实验素材是影响实验研究结果的核心因素之一, 它指的是心理学实验中使用的各种刺激材料, 实验素材的选择会直接影响实验的结果, 实验素材的选择需要考虑材料的性质及其容量。在认知行为实验研究中, 经常是同一种实验范式下相似的实验使用了非常不同的素材容量, 那么, 什么样的素材容量才是最合适的呢? 众所周知, 过少的实验素材会影响实验结果和结论的可靠性, 而过多的实验材料又可能导致被试疲劳及尾端实验数据不可靠, 同时还会导致实验耗时、增加实验成本等。因此, 选择一个最佳实验素材容量对实验结果就显得非常重要。

不同的实验素材容量明显会对实验信度产生

影响。在许多已有的实验研究文献中, 甚至同一类型实验, 实验素材的容量并不确定。如认知研究 IAT 范式中, 有采用 5 个(吴明证, 2006)、8 个(何安明, 刘华山, 惠秋平, 2013)等不同素材容量进行实验; 言语研究中图-词干扰范式上有采用 48 个(孙勇, 郭可教, 1992)、96 个(刘亚, 王振宏, 2011)、256 个(陈雪飞, 蒋军, 赵晓, 陈安涛, 2010)等不同素材容量; 注意研究的线索提示范式中, 有研究者选择了 160 个(潘运, 沈德立, 王杰, 2009)、256 个(李迎娣, 张学民, 2010)等不同素材容量; 在双耳分听范式(张学民, 舒华, 唐云, 2008; 郭春彦, 朱谨, 侯培庄, 1998)与情绪启动效应(杨丽珠, 蒋重清, 刘颖, 2005; 吕勇, 张伟娜, 沈德立, 2010)等研究范式中也存在这种现象。由此可以看到, 不同研究在相同类型范式的研究中采用的素材容量都存在着明显的不同, 虽然这些具体的实验之间存在着差异, 但是, 素材容量的选择还是存在着许多的主观性, 因此, 探讨最佳实验素材容量的选择, 从而更好的保证实验结果和结论的信度就显得非常重要。然而, 至今未发现专门对认知行为实验中最佳素材容量问题进行过专门的研究。

概化理论(Generalizability Theory, GT) 是一种把测量误差作为模型参数来处理的测量理论(Brennan, 2001; 杨志明, 张雷, 2003), 通过分析各种测量误差来源对测量目标的影响模式, 进而提出改进研究方案, 降低实验误差, 从而最终提高测量结果概化能力(Generalizability)的最佳方案。概化理论中关于变量的关键概念是测量目标和测量侧面, 测量目标一般就是被试在某项任务上的成绩, 而测量侧面就是影响测量目标的变量, 也就是测量误差来源。概化理论分为一元与多元概化理论。多元概化理论(Multivariate Generalizability Theory, MGT) 是对一元概化理论(Univariate Generalizability Theory, UGT) 的推广与发展。多元概化理论指测量目标在某个特定全域分数上有多个全域分数。当前, 针对 MGT 开发的常用软件是 mGENOVA (Brennan, 2001), 该软件使用比较简单, 即使对不太熟悉概化理论的人也容易学会。其分析结果一般包括(杨志明, 张雷, 2003): 第一, 分别估计测量目标、测量侧面以及目标与侧面或侧面与侧面之间的主效应和交互效应在各个变量的方差与变量间协方差; 第二, 估计考生在各个变量上全域分数、全域分数的方差、绝对误差、相对误差等, 并估计出测量目标在各个变量上的概化系数和可靠性指数; 第三, 估计全域合成分数及其方差; 第四, 估计全域合成分数的绝对误差方差和相对误差方差以及全域合成分数的可靠性指数、概化系数和信噪比; 第五, 对研究者所提出的改进方案估计其合理性。

概化理论根据样本值估计总体真值, 同时通过 D 研究指出真值所适用的空间与概括全域以及相应的精度估计值(G 系数与 ϕ 指数), 并且以相对误差与绝对误差的形式探讨了经典测量理论中的信度与效度。概化理论可以在多个范围上对测验分数做推论与解释, 并且分别提供测量的精度指标, 即一个测验提供多个“信度”值, 所以概化理论被广泛的应用在考试(杨志明, 张雷, 马世晔, 2004; 白娟, 2013)、量表编制(杨志明, 张雷, 2003; 何立国, 周爱保, 2006)与人才测评(孙晓敏, 张厚粲, 薛刚, 黎坚, 2009; 康春花, 姜宇, 辛涛, 2010)等研究中, 探讨不同变量的容量与整体信度(概化系数与可靠性指数)之间的关系。

本研究的目的是, 通过多元概化理论分析技术对 3 个经典认知行为实验范式下的具体实验数据进行分析, 来研究认知行为实验研究中最佳实验素材容量的确定问题。以为其他不同的实验研究在素材

容量选择上提供借鉴, 在保证实验信度的前提下, 使得实验设计更加科学合理。

2 研究方法

本研究选择了 3 个经典认知行为实验研究范式的具体实验, 即, 研究认知的 IAT 范式(Greenwald, McGhee, & Schwartz, 1998)、言语的图-词干扰范式(Stroop, 1935)与注意的线索提示范式(Posner, 1980), 实验程序统一采用 E-prime 1.1 软件编制, 数据处理基于多元概化理论, 运用 mGENOVA 软件对 3 个实验的数据进行分析, 以保证实验信度合理为条件, 选择和确定各实验中最合适的素材容量。

实验设计中采用的因变量均是反应时, 在进行多元概化理论分析前, 3 个实验所收集的数据都要经过筛选与预处理, 统一要求每个被试的正确率在 80%以上, 并且每个素材所有被试在该项目上的作答正确率为 80%以上, 否则作为无效被试或素材, 在此基础上, 对于作答错误素材, 采用在该素材上作答正确被试的平均反应时替换。

3 个认知行为实验采用的是相同的数据分析模型, 其中 G 研究包括两个维度的单面 $p \times i$ 交叉设计, X_{pi} 为被试在维度 h ($h = 1, 2$) 中第 i 个实验材料上的正确反应时, 则 G 研究的数学模型为:

$$X_{pi} = u + V_i + V_p + V_{pi} \quad (1)$$

在模型(1)中, p 是指测量目标, i 指测量的侧面, pi 为测量目标与测量侧面的交互效应。

2.1 研究 1 实验素材容量选择: IAT 范式 MGT 分析

2.1.1 IAT 范式研究过程

被试选择: 从某校选择 40 名大一新生做被试, 其中男生 20 人, 女生 20 人, 所有被试都是右利手, 矫正视力正常, 没有做过类似实验。

实验程序: 实验材料选择了黄色和红色作为彩色概念, 黑色和灰色做为非彩色概念, 同时选择了褒、贬词各 10 个高频词汇; 仿照 Greenwald 花虫实验七步骤程序流程, 只选取第四步与第七步的数据, 两步素材容量分别为 20, 共 40 个刺激, 注视点(黑色加号+)呈现时间为 500 ms, 被试按键反应后或者 4000 ms 未做出反应, 则刺激消失。实验流程通过 e-prime 1.1 软件编制程序, 以保证各个被试接收的刺激同等, 实验实施统一在一个光线适当的机房进行, 由负责人统一开始, 旁边由经过专门指导的人员负责被试在实验过程中可能遇到的问题。

数据预处理先按照统一要求筛选合格数据, 再

对高于 3000 ms 或者低于 300 ms 的数据分别替换为 3000 ms 和 300 ms, 之后替换作答错误素材的反应时。对于是否要求对数据进行自然对数转换, 目前研究尚有争议, 本文为了与后续研究一致, 统一不进行转换。

2.1.2 IAT 范式 G 研究

基于模型(1)通过软件 mGENOVA, 可以得到被试(p)、任务(i: 相容与不相容任务)以及被试与任务(pi)的交互效应在相容与不相容任务上的方差与协方差分量的估计结果, 如表 1。

表 1 G 研究中效应在不同任务上的方差与协方差分量估计

效应	相容任务	不相容任务
p	33140.42	0.40
	8531.42	13842.32
i	1672.28	16983.83
	75145.31	203154.13
pi		

注: 各效应主对角线上的元素为各效应在相应任务上的方差分量估计, 效应 p 主对角线以下元素为不同任务间协方差分量的估计, 主对角线之上元素为任务间的相关系数的估计。

对 IAT 的概化理论分析是基于 IAT 效应存在的前提下进行的, 从相容与不相容任务的平均反应时发现, $IAT\ 效应 = M_{不相容} - M_{相容} = 760.76 - 653.33 = 107.43$, 说明本实验存在 IAT 效应并且能够作为 IAT 范式的一个代表样例。由表 1 可知, 在 p (被试) 主效应上, 相容任务的方差分量要比不相容任务的方差分量大, 这说明被试在相容任务上的变异要大于不相容任务。同时从协方差分量与相关系数可以看出, 相容任务与不相容任务的相关比较低, 相关系数只有 0.40; 在 i (项目) 主效应上, 不相容任务的方差分量要高于相容任务的方差分量, 说明不相容任务的项目引起的变异要高于相容任务; 在 pi 交互效应上, 不相容任务的方差分量也要大于相容任务的方差分量。

2.1.3 IAT 范式 D 研究

根据 G 研究估计的方差和协方差矩阵, 可以进一步估计不同任务上的全域分数、误差的方差分量、概化系数及可靠性指数, 结果见表 2。

从表 2 可以看出, 相容任务的全域方差分量(33140.42)要小于不相容任务的全域方差分量(13842.32)。从测量精度来说, 由于本文是实验研究,

不是常模参照测验或标准参照测验, 必须要考虑测量的绝对误差, 在可靠性指数上, 不相容任务为 0.56, 明显要低于相容任务(0.90), 而从绝对误差上来看, 相容任务要低于不相容任务的方差分量也能反应出来。

在内隐联想实验中, 相容与不相容任务的实验素材容量是一样的, 所以两者的权重系数相同, 各为 0.5, 在此基础上合成全域分数, 结果如表 3。

表 2 不同任务上的方差分量值

指标名称	相容任务	不相容任务
全域分数	33140.42	13842.32
相对误差	3757.27	10157.71
绝对误差	3840.88	11006.90
均值误差	1006.06	1449.19
概化系数	0.90	0.58
可靠性指数	0.90	0.56

表 3 D 研究合成全域分数的方差分量等指标估计

指标	估计值
全域总分方差分量	16011.39
全域总分相对误差的方差分量	3478.74
全域总分绝对误差的方差分量	3711.94
全域总分误差均值的方差分量	720.46
全域总分的标准差(SD)	126.54
全域总分相对误差的标准差(SD)	58.98
全域总分绝对误差的标准差(SD)	60.93
全域总分误差均值的标准差(SD)	26.84
全域总分的概化系数	0.82
全域总分的可靠性指数	0.81
全域总分的相对信噪比	4.60
全域总分的绝对信噪比	4.31

从表 3 中数据可以看出, 全域概化系数值为 0.82, 可靠性指数为 0.81, 处于中等水平。全域总分相对误差的方差分量(3478.74)与绝对误差的方差分量(3711.94)差异较小, 所以概化系数与可靠性指数差异不大, 同时从可靠性指数大小而言是可以接受的, 测量精度中等。

实验分析了不同任务类型对总方差的贡献比例, 结果如下表 4。

表 4 不同任务对总方差贡献比例

方差贡献量	相容任务	不相容任务
全域总分 SD	65.07%	34.93%
相对误差 SD	27.00%	73.00%
绝对误差 SD	25.87%	74.13%

不同任务素材容量一致,从理论上说,按照任务均等影响,其贡献比例应是相同,但受到 p (被试)、i (任务)以及 pi 交互效应的影响,导致相容与不相容任务的贡献比例出现变化。在表 4 中可以看出, p 主效应在相容任务的方差分量大于不相容任务的方差分量,相容任务对全域总分的方差贡献比例为 65.07%,而不相容任务只有 32.26%。在相对误差与绝对误差上,相容任务的贡献比例都低于不相容任务的贡献比例。

以保证 IAT 范式实验信度为目标,探讨最佳的实验素材容量的选择,结果见表 5。

表 5 D 研究中不同任务样本量与可靠性指数关系

	n`1	n`2	总分可 靠性指数	可靠性 指数变化
G 研究样本	20	20	0.79	
D 研究样本				
基准模型	1	1	0.18	
对比模式 1	5	5	0.52	+0.34
0.5 倍模式	10	10	0.68	+0.16
G 样本模式	20	20	0.81	+0.13
1.5 倍模式	30	30	0.87	+0.06
2 倍模式	40	40	0.90	+0.03
2.5 倍模式	50	50	0.92	+0.02
3 倍模式	60	60	0.93	+0.01
3.5 倍模式	70	70	0.94	+0.01
4 倍模式	80	80	0.95	+0.01

当不同任务的素材容量只有 1 个时(基准),总分可靠性指数只有 0.18,素材容量增至 5 时,总分可靠性指数也只有 0.52,增加了 0.34,在实验中,这样的精度是不被接受的,于是素材容量增加到 10,也就是不同任务各有 10 个,总分可靠性指数增加了 0.16,达到了 0.68,在实际中,这个实验信度还是偏低,于是当增加到 20,也就是本文中 G 研究的样本素材容量时,可靠性指数增加了 0.13,达到了中等精度要求 0.81,之后继续增加实验素材容量,在增加到 60、70、80 时,增量之间的差异非常小,趋于稳定,相对在 2.5 倍模式时,即素材容量为 50 时,增量的差异为 0.01,而且增量为 0.02,是随着素材容量增加,可靠性指数变化趋势的转折点,同时全域总分可靠性指数为 0.92,综合认为在 IAT 范式中,最佳的素材容量为 50。

2.2 研究 2 实验素材容量选择: 图-词干扰范式 (stroop 效应)MGT 分析

2.2.1 图-词干扰范式研究过程

被试选择: 从某校选择 31 名大二新生,其中男

生 10 人,女生 21 人,所有被试都是右利手,矫正视力正常,没有做过类似实验。

实验程序: 图片采用 250×200 像素,字体大小为 100 磅,干扰条件用红、绿、黄、蓝 4 种颜色书写红、绿、黄、蓝汉字(不包括呈现颜色与内容颜色一致的材料),控制任务用红、绿、黄、蓝 4 种颜色正三角形、圆形和正方形、正六边形 4 种几何图形,每个刺激在实验中重复呈现 2 次,并且从 4 个图形中抽取一种颜色作为练习,使得干扰与控制素材容量一致,所以被试在正式实验中,控制与色词素材容量分别为 24 个,实验要求忽略词义对呈现的颜色作按键反应,被试按键要求分别用左手中指、食指和右手的食指、中指按 D 键、F 键、J 键和 K 键(红、绿、黄、蓝),每个实验界面中下位置会出现 4 个颜色词与对应按键,注视点(黑色加号+)呈现时间为 500 ms,被试按键反应后或者 3000 ms 未做出反应,则刺激消失。实验采用 E-prime 程序编制,电脑统一采用 DELL 台式,在计算机机房统一进行。

数据预处理采用上文一致标准筛选,同时替换错误作答素材的反应时之后,再进行多元概化理论分析。

2.2.2 图-词干扰范式 G 研究

被试(p)、任务(i: 色词与控制任务)以及被试与任务(pi)的交互效应在色词与控制任务上的方差与协方差分量的估计结果,如表 6。

表 6 G 研究中效应在不同任务上的方差与协方差分量估计

效应	色词	控制
p	48249.66	0.96
	36845.63	30540.78
i	10767.89	4637.99
	189079.59	150154.69

注: 各效应主对角线上的元素为各效应在相应任务上的方差分量估计,效应 p 主对角线以下元素为不同任务间协方差分量的估计,主对角线之上元素为任务间的相关系数的估计。

由表 6 可以看出,首先保证 stroop 效应的存在,经过对色词与控制任务分别计算均值发现, stroop 效应 = M 色词 - M 控制 = 1092.58 - 966.04 = 126.54。在 p (被试)主效应上,色词的方差分量要高于控制任务的方差分量值;在测量侧面 i 上,控制任务的

方差分量为 4637.96, 而色词任务的方差分量为 10767.89, 所以控制任务的变异明显要低于色词任务; 同时在 π 交互效应上, 色词任务的方差分量也要高于控制任务, 从色词与控制任务的相关与协方差值可以看出, 两者呈高度相关。

根据 G 研究估计的方差和协方差矩阵, 可以进一步估计不同任务上的全域分数、误差的方差分量、概化系数及可靠性指数, 结果见表 7。

表 7 不同任务上的方差分量值

指标名称	色词	控制
全域分数	48249.66	30540.78
相对误差	7878.32	6256.45
绝对误差	8326.98	6449.70
均值误差	2259.24	1380.26
概化系数	0.86	0.83
可靠性指数	0.85	0.83

从全域分数的方差分量来看, 控制任务的方差分量明显要低于色词任务的方差分量。在概化系数与可靠性指数上, 控制任务与色词任务相差不大, 都为 0.82 以上, 达到了良好水平, 相对而言, 色词任务的测量精度更好一些, 但两者差异不明显。

2.2.3 图-词干扰范式 D 研究

实验中, 由于色词任务与控制任务的素材容量相同, 所以两者以 0.5 为权重合成了全域总分, 结果如表 8。

表 8 D 研究合成全域分数的方差分量等指标估计

指标	估计值
全域总分方差分量	38120.43
全域总分相对误差的方差分量	3533.69
全域总分的绝对误差的方差分量	3694.17
全域总分误差均值的方差分量	1504.16
全域总分的标准差(SD)	195.25
全域总分相对误差的标准差(SD)	59.45
全域总分绝对误差的标准差(SD)	60.78
全域总分误差均值的标准差(SD)	38.78
全域总分的概化系数	0.92
全域总分的可靠性指数	0.91

从表 8 可以看出, 全域总分的概化系数与可靠性指数都达到了 0.91 以上, 这是一个比较高的值, 说明图-词干扰范式的实验控制得比较好, 受误差影响较低, 精度较高。

不同任务对总方差贡献比例的分析结果如表 9

所示,

由表 9 可知, 色词任务的方差贡献比控制任务高 11%, 同时在相对误差与绝对误差上色词都要高于控制任务, 但各自在相对与绝对误差上差异较小。

以保证实验信度为目标, 探讨了本实验范式的最佳实验素材容量的选择, 结果见表 10。

表 9 不同任务对总方差贡献比例

方差贡献量	色词	控制
全域总分 SD	55.81%	44.19%
相对误差 SD	55.74%	44.26%
绝对误差 SD	56.35%	43.65%

表 10 D 研究中不同任务样本量与可靠性指数关系

	n'1	n'2	总分可靠性指数	可靠性指数变化
G 研究样本	24	24	0.91	
D 研究样本				
基准模型	1	1	0.30	
对比模式 1	6	6	0.72	+0.42
0.5 倍模式	12	12	0.84	+0.12
G 样本模式	24	24	0.91	+0.07
1.5 倍模式	36	36	0.94	+0.03
2 倍模式	48	48	0.95	+0.01
2.5 倍模式	60	60	0.96	+0.01

从表 10 可以看出, 色词与控制任务素材容量为基准模式时, 可靠性指数只有 0.30, 当容量达到 G 研究样本的 0.5 倍, 即 12 时, 可靠性指数就能达到 0.84, 并且比 0.25 倍容量增加了 0.12, 在增加一倍后, 即为 24 时, 增量为 0.07, 总分可靠性指数达到 0.91。当色词与控制任务素材容量各为 36 与 48 时, 增量分别为 0.03 和 0.01, 全域总分可靠性指数为 0.94 和 0.95, 之后继续增加, 可靠性指数的增量只有 0.01, 同时在 2 倍模式时, 与 1.5 倍模式增量变化为 0.01, 与 2.5 倍模式的增量比较也只有 0.01, 增量极低。综合考虑认为, 总分可靠性指数达到了 0.95, 素材容量为 48 时, 实验素材容量达到最佳, 继续增加素材容量时的可靠性指数变化不大。

2.3 研究 3 实验素材容量选择: 线索提示范式 MGT 分析

2.3.1 线索提示范式研究过程

实验被试: 实验同样收集了 31 个被试的反应时数据, 其中男生 10 人, 女生 21, 皆为大二学生。

实验程序: 实验图片都采用 240×170 像素, 以箭头为提示线索, “*”为目标刺激, 线索提示呈现时

间为 300~600 ms 之间随机一个点, 目标刺激呈现 1500 ms, 在实验前告知被试线索提示与目标一致性概率分别为 80%与 50%, 对左右不同位置的目标刺激分别按“d”键与“k”键。正式实验中, 有效提示与无效提示素材容量都为 25, 共 50 个素材。实验采用 E-prime 程序编制, 统一在学院机房进行。

数据预处理在线索提示范式实验中也采用上文一致标准筛选, 同时替换错误作答反应时。

2.3.2 线索提示范式 G 研究

被试(p)、任务(i: 有效提示与无效提示任务)以及被试与任务(pi)的交互效应在有效提示与无效提示任务上的方差与协方差分量的估计结果, 如表 11。

表 11 G 研究中效应在不同任务上的方差与协方差分量估计

效应	有效	无效
p	3214.10	0.82
	2894.60	3879.22
i	69.52	59.25
pi	6128.17	
		6381.67

注: 各效应主对角线上的元素为各效应在相应任务上的方差分量估计, 效应 p 主对角线以下元素为不同任务间协方差分量的估计, 主对角线之上元素为任务间的相关系数的估计。

在表 11 中, 首先对线索提示范式的效应存在时进行分析, 通过比较有效提示与无效提示均值发现, 提示效应 = $M_{\text{无效}} - M_{\text{有效}} = 414.35 - 377.42 = 36.93$, 说明提示效应是存在的。在 p 主效应上, 无效提示的方差分量(3879.22)要高于有效提示的方差分量(3214.10), 说明无效提示变异更大; 对于不同提示效果的测量侧面 i 可以看出, 两者的方差分量相差不大; 在 pi 交互效应上, 无效提示的方差分量要高于有效提示的方差分量, 同时从协方差与相关系数可以看出, 有效与无效任务的相关较高。

2.3.3 线索提示范式 D 研究

根据 G 研究估计的方差和协方差矩阵, 可以进一步估计不同任务上的全域分数、误差的方差分量、概化系数及可靠性指数。结果见表 12。

从表 12 可以看出, 在全域分数的分差分量中, 有效提示产生的方差分量明显要低于无效提示的方差分量。有效提示和无效提示的概化系数分别为 0.93 和 0.94, 可靠性指数也达到了 0.93 和 0.94, 说明实验的精度在有效与无效提示任务上都比较高,

受到误差影响较小。

由于有效提示为 50%, 所以有效提示与无效提示素材容量的权重是一样的, 从而合成了如表 13 所示的全域总分的不同数据指标。

表 12 不同任务上的方差分量值

指标名称	有效	无效
全域分数	3214.10	3879.22
相对误差	245.13	255.27
绝对误差	247.91	257.64
均值误差	114.37	135.74
概化系数	0.93	0.94
可靠性指数	0.93	0.94

表 13 D 研究合成全域分数的方差分量等指标估计

指标	估计值
全域总分方差分量	3220.63
全域总分相对误差的方差分量	125.10
全域总分的绝对误差的分差分量	126.39
全域总分误差均值的方差分量	109.22
全域总分的标准差(SD)	56.75
全域总分相对误差的标准差(SD)	11.19
全域总分绝对误差的标准差(SD)	11.24
全域总分误差均值的标准差(SD)	10.45
全域总分的概化系数	0.96
全域总分的可靠性指数	0.96
全域总分的相对信噪比	25.75
全域总分的绝对信噪比	25.48

全域总分的相对误差与绝对误差相差较小, 从全域总分的概化系数与可靠性指数的差异也可以看出来。全域总分的概化系数达到了 0.96, 而可靠性指数也为 0.96, 所以, 绝对与相对误差或者标准差之间的差异都比较小。全域总分的可靠性指数为 0.96, 这是非常高的一个数值, 说明线索提示范式的实验测量精度非常高。

不同提示任务对总方差的贡献比例分析结果, 如表 14 所示。

表 14 不同提示对总方差贡献比例

方差贡献量	有效	无效
全域总分 SD	47.42%	52.58%
相对误差 SD	48.99%	51.01%
绝对误差 SD	49.04%	50.96%

在表 14 中, 从不同提示任务对总方差的贡献比例可以看出, 虽然无效提示与有效提示的素材容

量是一样的,但是无效提示对总方差的贡献明显要高于有效提示的作用,两者分别为 52.58%、47.42%,但不同提示任务在相对误差与绝对误差上差别还是比较小的。

以保证实验信度为目标,探讨了本实验范式的最佳实验素材容量的选择,结果见表 15。

表 15 D 研究中不同提示样本量与可靠性指数关系

50%	n'1	n'2	总分可靠性指数	可靠性指数变化
G 研究样本	25	25	0.96	
D 研究样本				
基准模型	1	1	0.51	
0.5 倍模式	15	15	0.94	+0.43
G 样本模式	25	25	0.96	+0.02
1.5 倍模式	35	35	0.97	+0.01
2 倍模式	50	50	0.98	+0.01
2.5 倍模式	60	60	0.98	+0.00

注:对于 0.5、1.5、2.5 倍模式,考虑到素材容量整数变化趋势,进行了些许调整。

在表 15 中,基准模式的可靠性指数只有 0.51,到 G 研究样本容量 25 时,可靠性指数达到了 0.96 以上,相对 0.5 倍模式,增加了 0.02,当素材容量增加至 35 时,增量为 0.01,可靠性指数为 0.97,当达到 2 倍模式与 2.5 倍模式时,增量分别为 0.01、0.00,可靠性指数达到了 0.98、0.98,但可靠性指数增量明显变小,所以综合而言,线索提示范式的最佳素材容量为 35,可靠性指数为 0.97。

3 讨论与结论

在许多的认知行为实验研究中,典型的素材容量的选择从 20 左右到 100 左右,相互相差很大,还

未发现研究者对各种容量选择情形下的误差进行解释。那么,当容量为 20 时就够了吗?达到 100 时是否合适呢?

在认知行为实验研究中,如何确定最佳的实验素材容量?有的研究者可能依据前人的研究设计,有的研究者可能依据自身的经验,有的研究者可能会依据认知行为实验研究的一般性要求,有的研究者可能只能根据素材的实际可得性条件来确定。许多相同的实验范式之间使用的素材容量相差极大。素材容量主要是影响实验结果的信度,即结果的稳定性、一致性程度。研究者当然可以通过不断增加容量以确保实验结果的稳定性,然而,究竟多少是合适的呢?研究者也许可以通过不断试验的方式进行研究,但是,基于概化理论的分析也许是一种更加直接高效的分析架构,因为它既可以分析出一种理想的素材容量,同时也可以指出当素材无法达到最佳容量要求时可能存在的误差大小。概化理论可以在同时考虑多个影响测量目标的因素的情形下分析出最佳的实验设计模式。

本文运用多元概化理论技术,以追求良好的实验信度为目标,同时考虑实验成本,探讨了不同实验(认知、语言、注意)的素材容量的最佳选择。同时在确定实验效应存在的前提下,通过素材容量的改变,结合可靠性指数的增量变化大小,从而确定最佳实验素材容量,为实验研究在素材容量选择上提供有价值的参考。结合前面对 IAT 内隐联想范式、图-词干扰范式、线索提示范式的分析,以及图 1 所示,可以认为,在 IAT 实验上,相容与不相容任务的素材容量为 50 时,即 2.5 倍模式,可靠性指数为 0.92,之后实验的可靠性指数变化趋于直线;在图-词干扰范式中,素材容量位于 2 倍模式后,增幅趋

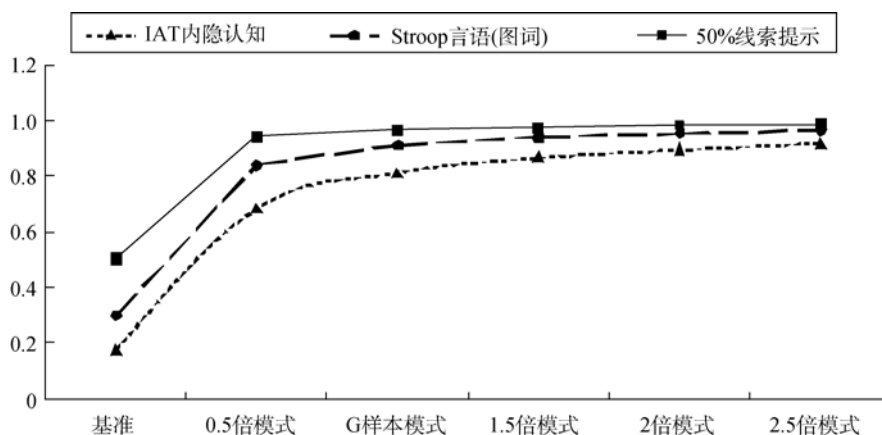


图 1 不同实验素材容量的可靠性指数

于直线, 所以其素材容量以 48 为最佳, 可靠性指数为 0.95; 对线索提示范式, 当 50% 线索为有效提示时, 有效与无效提示最佳素材容量为 35, 即素材容量为 1.5 倍模式, 其测量精度为 0.97。

本研究可以为认知行为实验研究者在选择合适的实验素材容量时提供一些有益的借鉴。

参 考 文 献

- Bai, J. (2003). Study of the multivariate generalizability analysis of Chinese medicine entrance tests in 2012. *Examinations Research*, (1), 51–56.
- [白娟. (2013). 2012 年全国硕士研究生入学中医综合考试的多元概化理论研究. *考试研究*, (1), 51–56.]
- Brennan, R. L. (2001). *Generalizability theory*. New York: Springer-Verlag.
- Chen, X. F., Jiang, J., Zhao, X., & Chen, A. T. (2010). Effects of practice on semantic conflict and response conflict in the Stroop task. *Psychological Science*, 33(4), 869–871.
- [陈雪飞, 蒋军, 赵晓, 陈安涛. (2010). 练习对 Stroop 效应中语义冲突和反应冲突的影响. *心理科学*, 33(4), 869–871.]
- Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personal and Social Psychology*, 74(6), 1462–1480.
- Guo, C. Y., Zhu, Y., & Hou, P. Z. (1998). An experimental study of the influence of binaural hearing and word-fragment completion on immediate recall and delayed recall. *Psychological Science*, 21(2), 118–122.
- [郭春彦, 朱滢, 侯培庄. (1998). 双耳分听和补笔对立即回忆、延迟回忆影响的实验研究. *心理科学*, 21(2), 118–122.]
- He, A. M., Liu, H. S., & Hui, Q. P. (2013). An experimental research on implicit and explicit gratitude of undergraduates. *Psychological Development and Education*, 45(1), 22–30.
- [何安明, 刘华山, 惠秋平. (2013). 大学生感恩内隐效应的实验研究. *心理发展与教育*, 45(1), 22–30.]
- He, L. G., & Zhou, A. B. (2006). A study of life satisfaction scales applicable to Chinese adolescent students by the Generalizability theory. *Psychological Science*, 29(5), 1199–1202.
- [何立国, 周爱保. (2006). “青少年学生生活满意度量表”的概化理论研究. *心理科学*, 29(5), 1199–1202.]
- Kang, C. H., Jiang, Y., & Xin, T. (2010). Generalizability theory in the study of raters' consistency of personnel evaluation. *Psychological Science*, 33(6), 1456–1460.
- [康春花, 姜宇, 辛涛. (2010). 概化理论在人事测评中的评分者一致性研究. *心理科学*, 33(6), 1456–1460.]
- Li, Y. D., & Zhang, X. M. (2010). The different cuing effect of three types of cue. *Psychological Science*, 33(3), 627–630.
- [李迎娣, 张学民. (2010). 不同类型提示线索提示效应的大小比较. *心理科学*, 33(3), 627–630.]
- Liu, Y., & Wang, Z. H. (2011). The relationship between the emotional Stroop effect and the Stroop effect. *Psychological Science*, 34(4), 806–812.
- [刘亚, 王振宏. (2011). 情绪 Stroop 效应与 Stroop 效应的关系. *心理科学*, 34(4), 806–812.]
- Lü, Y., Zhang, W. N., & Shen, D. L. (2010). Subliminal affective priming effect by faces with different valence: An ERP study. *Acta Psychologica Sinica*, 42(9), 929–938.
- [吕勇, 张伟娜, 沈德立. (2010). 不同愉悦度面孔阈下情绪启动效应, 来自 ERP 的证据. *心理学报*, 42(9), 929–938.]
- Pan, Y., Shen, D. L., & Wang, J. (2009). The spatial numerical association of response codes effect of Chinese character number processing in different attention referential cues. *Studies of Psychology and Behavior*, 7(1), 21–26.
- [潘运, 沈德立, 王杰. (2009). 不同注意提示线索条件下汉字数字加工的 SNARC 效应. *心理与行为研究*, 7(1), 21–26.]
- Posner, M. I. (1980). Orienting of attention. *Quarterly Journal of Experimental Psychology*, 32, 22–25.
- Stroop, J. R. (1935). Studies of interference in serial-verbal reaction. *Journal of Experimental Psychology*, 18(6), 643–662.
- Sun, X. M., Zhang, H. C., Xue, G., & Li, J. (2009). The application of the multivariate generalizability theory to a structured interview. *Psychological Science*, 32(4), 916–919.
- [孙晓敏, 张厚粲, 薛刚, 黎坚. (2009). 多元概化理论在结构化面试中的应用研究. *心理科学*, 32(4), 916–919.]
- Sun, Y., & Guo, K. J. (1992). Hemispheric asymmetry in the processing of Stroop stimuli. *Psychological Science*, (2), 20–25.
- [孙勇, 郭可教. (1992). Stroop 色词干扰效应与大脑两半球言语功能的不对称性. *心理科学*, (2), 20–25.]
- Wu, M. Z. (2006). The relationship between the IAT effect and the target-concept's association. *Psychological Science*, 29(3), 580–582, 531.
- [吴明证. (2006). 内隐联结测验的 IAT 效应和概念联系的对立关系研究. *心理科学*, 29(3), 580–582, 531.]
- Yang, L. Z., Jiang, Z. Q., & Liu, Y. (2005). A comparative study of subliminal affective priming and the emotional Stroop effect. *Psychological Science*, 28(4), 784–787.
- [杨丽珠, 蒋重清, 刘颖. (2005). 阈下情绪启动效应和 Stroop 效应之对比实验研究. *心理科学*, 28(4), 784–787.]
- Yang Z. M., & Zhang, L. (2003). *Generalizability theory and its applications*. Beijing: Educational Science Publishing House.
- [杨志明, 张雷. (2003). *测评的概化理论及其应用*. 北京: 教育科学出版社.]
- Yang, Z. M., & Zhang, L. (2003). Can we measure the third factor by WISC-CR? —a study of WISC CR by the Multivariate Generalizability Theory. *Psychological Science*, 26(2), 305–307.
- [杨志明, 张雷. (2003). 韦氏儿童智力量表能否测量第 3 因子—WISC-CR 的多元概化理论研究. *心理科学*, 26(2), 305–307.]
- Yang, Z. M., Zhang, L., & Ma, S. Y. (2004). Multivariate generalizability analysis of the Chinese college entrance comprehensive examination. *Psychological Science*, 36(2), 195–200.
- [杨志明, 张雷, 马世晔. (2004). 从多元概化理论看高考综合能力测试的改进. *心理学报*, 36(2), 195–200.]
- Zhang, X. M., Shu, H., & Tang, Y. (2008). Phonological effect on Chinese word cognition in dichotic listening task. *Applied Linguistics*, (2), 72–80.
- [张学民, 舒华, 唐云. (2008). 双耳分听任务中语音相似性对汉字认知的影响. *语言文字应用*, (2), 72–80.]

The Optimal Size of Material in Psychological Experiment: The Applications of Multivariate Generalizability Theory

LUO Zhaosheng; GUO Xiaojun

(School of psychology, Jiangxi Normal University, Nanchang 330022, China)

Abstract

The validity and reliability must be guaranteed to any psychological experiment. The size of experiment material is a critical factor that will significantly affect the results of experiment. This paper aims to study the selection of optimal size of the psychological experiment material under the frame of multivariate generalizability theory. The results show that, with the IAT experiment, 50 will be the optimal size of each combined tasks, and the reliability will be 0.905; with the Picture-word Interference Paradigm, when the material size is about 48, the reliability will be 0.954; with the Cue Paradigm, when the number of valid cue is 50%, the optimum size of valid cue and invalid cue is about 35, and the reliability will be 0.973. This study shows that MGT can help us to select an optimal size of the psychological experiment material.

Key words multivariate generalizability theory; the reliability of experiment; size of experiment material