

让数据站住脚-浅谈用户研究中的信度与效度

在用户研究工作中，如何让自己的数据和结论更有说服力，是很重要的问题。最近将自己积累的用研信度和效度的笔记整理一下，罗列在文中，希望对大家有所帮助。

一、调查的质量取决于调查的信度和效度

信度主要指测量结果的一致性、稳定性。也就是说结论和数据是否反映了用户最真实稳定的想法。用户在回答问题的时候，往往会受到环境、时间、当时当地的情绪影响，而作出并不真实的想法，即会有随机误差。信度就是衡量这种随机误差对用户想法的影响大小。

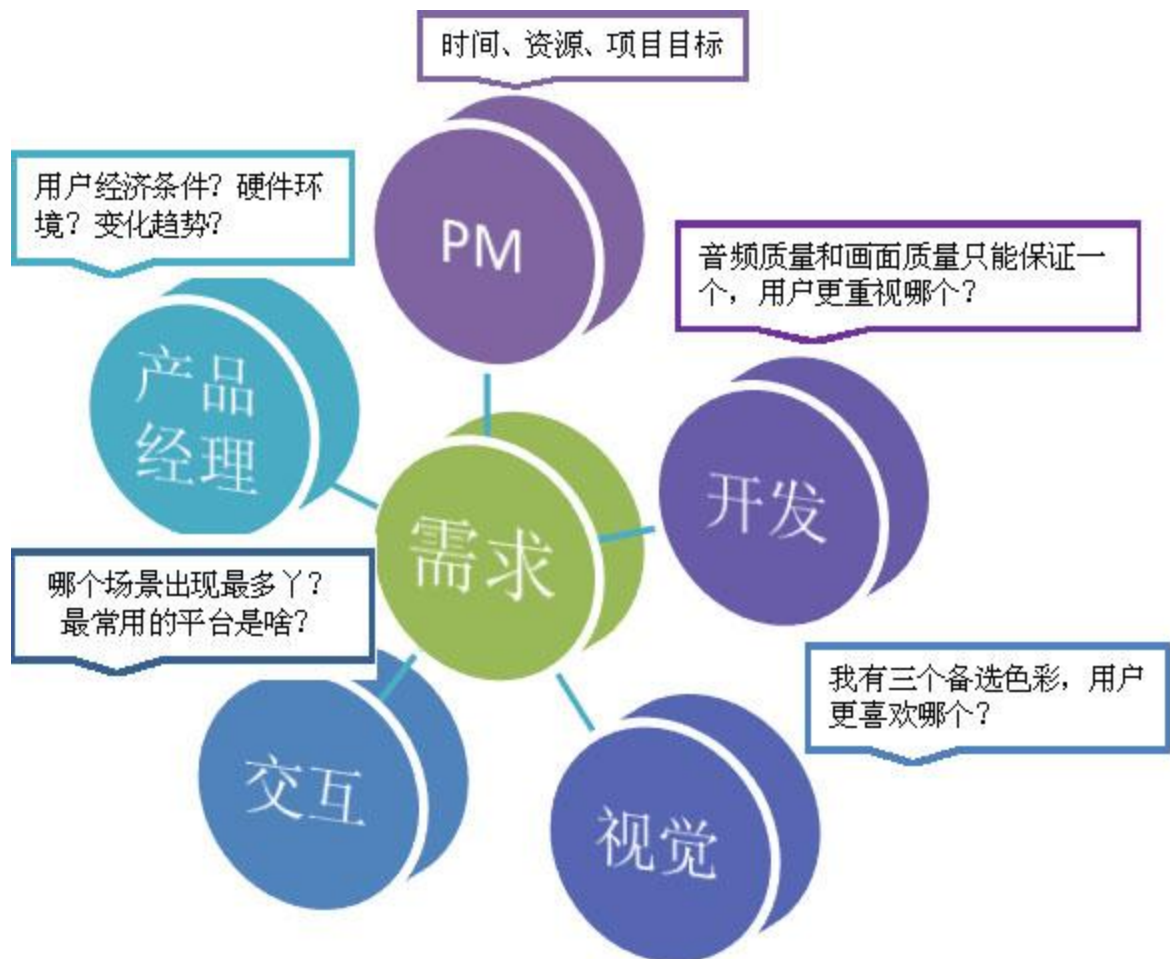
效度是指多大程度上测量了你想要测量的东西。

对某个产品用研，我们现在用得最多是用户访谈、问卷调查和可用性测试。而在这几个过程中都会涉及信度和效度的问题。

二、用户访谈中的效度和信度

1. 访谈不能仅仅局限于用户

任何一个产品项目都会受到市场环境、公司战略、技术力量、平台规范和流行趋势等各个方面的影响。对某一产品的需求，可能来自用户、产品、技术、交互以及视觉。不同岗位人员看待产品的角度不一样，侧重点也不一样，找多个角色有助于把需求找全，不遗漏，所以必须提前了解他们的需求。这样才能使我们的研究更有针对性、全面性、有用性。有用程度、全面程度是效度的重要组成部分。



2. 巧妙的选择访谈用户

通常，前期深度访谈的用户数量不会太多，所以用户条件一定要把握适当。**反馈的问题才能全面、合理、有用。**

比如是做 Android 平台上的某一软件。

首先 **Android 新用户和熟练用户都是必须的**，熟练用户更能反映 android 用户习惯性操作方式、平台特点、以及长期使用过程中积累的意见和建议；而新手用户可以更好的反映该平台哪些地方存在学习困难，从而通过我们的设计帮助用户去降低学习成本。

其次**非 Android 平台用户也是必须的**，可以从侧面了解他们不用 Android 的原因。从而帮助产品挖掘更多潜在用户提供方向。

人口学信息（学历、职业、性别、年龄）要覆盖全面。不同属性的用户看重地方会存在差异。需求也会不一样。

包含竞品用户。通过了解用户对竞品的评价，可以提炼出竞品的优劣势，从而为增强产品竞争力提供方向。

3. 一定要有专家

专家是重要的信息携带者。李乐山教授说专家有三类，用户专家、制造专家、市场销售专家，他指出判断某人是否是专家的标准是：（1）能够熟练使用一种产品；（2）能够比较同类产品；（3）有关的新知识容易整合到自己的知识结构中；（4）具有 10 年专业经验；（5）积累大量经验并且在使用经验方面具有绝招；（6）了解有关的历史（该产品设计史、技术发展史等）；（7）关注产品发展趋势；（8）知识链或者思维链 比较长，提起任何一个有关话题，他们都能够谈出大量的有关信息；（9）能够提出改进或创新的建议，他们的创新或改进方案，其高水平体现在采用简单方法解决 复杂问题。

对于互联网，专家应该指的是用户专家、开发专家、设计专家以及产品专家；他们凭借丰富的经验，系统全面的掌握行业同类产品、开发及设计模式、历史及发展趋势、专业水平极高。他们可以为我们提供很多我们始料未及的建议。这是保证用研过程，特别是对于后期问卷结构效度有很大的作用。

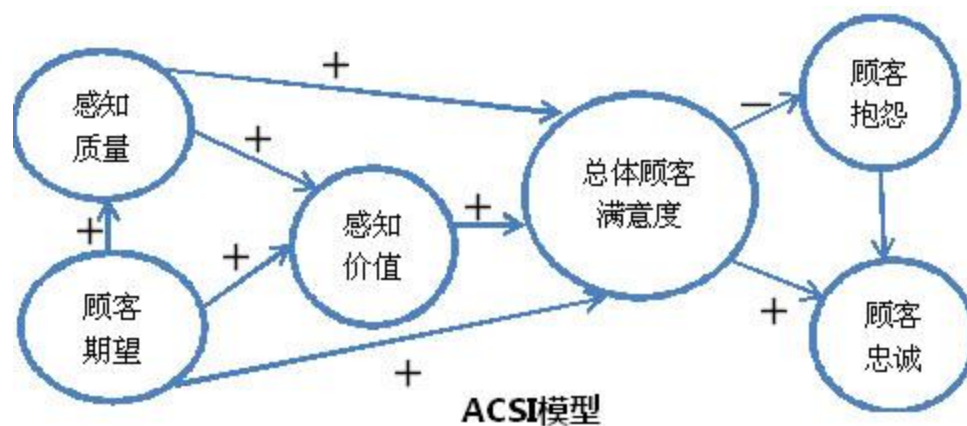
三、问卷调查与分析中的信度与效度

为了提高工作效率，问卷调查往往采用网络调查的方法，信度效度问题出现的可能性就更大。

最近看到一些满意度调查是采用量表加结构方程模型（SEM）的方式。我们看看哪些地方可能会出现信度和效度的问题。

1. 理论模型支持

由于 SEM 进行的是验证性因子分析，是检验而不是探索新的模型，因此，整个因果关系的假设必须有强有力的理论支持和严密的逻辑框架。包括模型中 变量关系的假定、指标的选取、甚至测度项的表达方式等。如果最终输出的模型和理论模型结构不符，那么该模型是没有任何说服力的。比如用 ACSI 模型作为满意度的理论模型时，是否真的按照感知质量、感知价值、顾客期望这几个层面去设计问卷？



2. 保证份量

普通抽样调查中原则上是越多越好，但遇到目标用户较少的情况，只要保证一定的条件就 ok 的，样本量受到置信区间、抽样误差范围的影响，可以用公式算出最小样本量。

但对于结构方程模型大样本是必须的，SEM 中涉及的变量众多，变量间的关系很复杂交错，小样本量会导致模型不稳定，收敛失败进而影响模型中参数。朱远程等^[1]在文献中指出，当样本低于 100 时，几乎所有的结构方程模型分析都是不稳定的，大于 200 以上的样本，才称得上一个中型样本。若要得到稳定的结构方程模型结构，低于 200 的样本数量是不鼓励的。有些学者将最低样本量与模型变量结合在一起，建议样本数至少应为变量的十倍，这一规则经常被引用。模型中变量越多，对大样本的要求就越高。

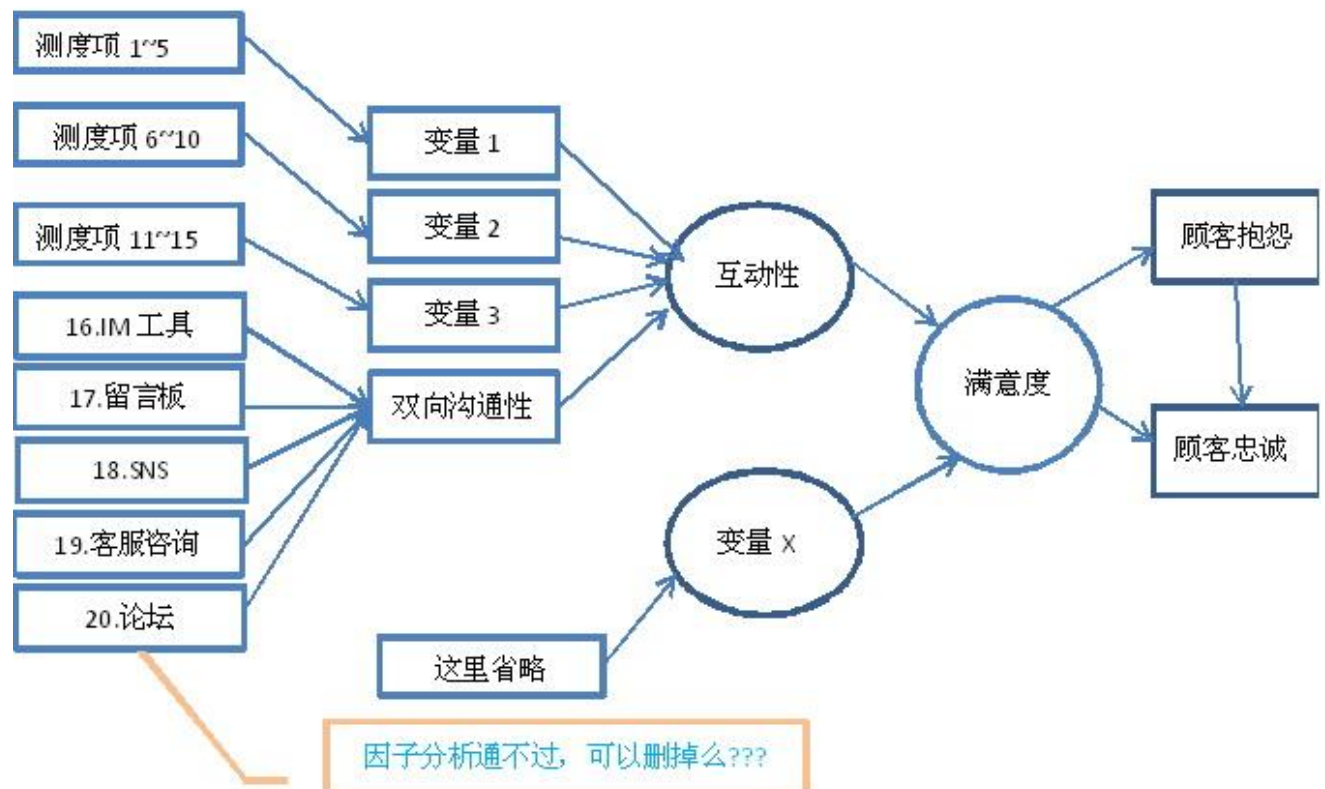


3. 变量需遵循原则

- SEM 模型中各变量的函数关系要是线性的，否则是不能用回归计算路径系数的。
- 在使用最大似然估计法时，变量一定要是多元正态分布的，这就要求指标要呈正态分布，否则就要对指标进行正态处理才行。
- 变量间的多重共线性程度要低，否则路径系数会有很大误差。
- SEM 建立的过程中会不断的修正才能得到比较完美的模型，比如因子分析时，若发现某一测度项对应的因子载荷过小，就会人为的将该测度项删除，但是若模型建立之后，一些变量对应了 4~5 个测度项，一些变量只剩下 1~2 个测度项，那么我们就需要思考只有两个测度项的变量是否被完全解释，这仅有的两个测度项就全面真实的反映该变量么？如果是这样，就算 KMO、Bartlett、因子载荷都通过了，效度也是难以保证的。所以问卷前期需要反复的预调研，不断的对问题进行修正，而不是随意的人为删除。我学生时代对某电子商务网站满意度进行调查时，就犯了类似的错误，模型中的“互动性”片段，互动性由四个变量

衡量，其中“双向沟通性”一开始设计的时候由 5 个测度项支持，但是因子分析检验通不过，就直接将因子载荷比较小的客服、论坛、SNS 三者去掉了，最后虽然在数据上通过了信度效度检验，但是只有 IM、留言板这两个测度项支持是绝对不能解释“双向沟通性”的。

某电子商务网站满意度模型部分框架



4. 数据质量是根源

要使模型结构稳定有效，首先要保证数据质量，反复检验问卷的信度。

a. 不同时间的一致性。

在设计问卷时，可以将同样的问题对同一个人重复测试，如果这两道题得到的答案是不一致的，相关系数（Pearson r ）小于 0.7，那么这份问卷的稳定信度就值得考量。

假如问卷样本足够大，可以一分为二（每一个样本也要保证足够样本量），分别建立两个模型；通过对比两个模型中参数的差异，便可以检验该模型的稳定性和适用性。如果两者差异太大，就说明模型本身是有问题的。

咦~~咱俩是一个妈生的么



b. 不同形式的一致性

用内容等效但表达方式不同的两份问卷调查，检测两者的等效信度，比如 **Gamma** 系数。

c. 内在一致性

问卷中相关的问题为同样的目标服务，他们在逻辑一致，也就是同质的。首先要测量每个测度项与总体的相关性（**item-total correlation**），然后再测量同一变量下相关问题间的同质性，而对于不同的提问方式选择对应的方法：比如,对于李克特量表方法，就用 **Chronbach** 系数检验；在基础研究中，信度至少应达到 **0.80** 才可接受，在探索性研究中，**0.70** 可接受，**0.70—0.98** 为高信度,小于 **0.35** 为低信度。对于是非题则采用 **kuder-Richardson** 系数检验。在进行内在一致性检验时，要看题目选项是否反序，如果两道题都是问“对该产品是否 满意”，一道 **7** 代表满意，**1** 代表不满意；另一道 **1** 代表满意，**7** 代表不满意，这样就会影响信度。遇到这种情况要提前人为调整过来。

5. 看得更远一点

问卷结论不仅要解决当前的问题和需求，还有具有一定的预测作用，市场是变化的，当前的目标用户不一定是未来的（或者下一个版本的）目标用户，比如目标用户的收入可能有增加的趋势，某一平台的使用率在快速提高，当前的满意度模型可能在一个月之后就不适用了（比如新功能点的出现）。



假设我们要对 **QQ** 影音进行满意度调查，现在建立了一个满意度模型，但若下个月 **QQ** 影音中多了一个重要的功能，对整个满意度的提升产生了很大作用，那么，模型中各项的路径系数会不会产生变化？该模型在下个月可能就不适用了，造成的后果就是当前的满意度值与下个月的满意度值没有可比性了，很多工作也就白费了。所以，诸如满意度模型这样的研究，是需要反复调查，长期对该满意度模型进行监控和修正，以求得到最稳定的模型，就可以让模型会具有很预测和比对作用啦。

6.关注细节

- a. 问卷设计中题项表述不能出现歧义、避免太专业词汇以及诱导词汇
- b. 选项间要有明确的区分（互斥）
- c. 避免遗漏，“其他”选项是必须的，而且最好配有输入框，记忆中，每次问卷调查中都能从“其他”选项中获取大量信息。
- d. 一般题项不能太多，设置问题选项的时候，尽可能的让选项**随机显示**，特别是在选项较多的情况下。
- e. 数据处理过程中删除重复项矛盾项之外，最好能统计到用户**填写问卷的时间差**。如果整个填写的时间极短，完全可以判定用户没有认真填写。
- f. 极端的、离群的选项可以考虑将其删除。

四、可用性测试中的信度与效度

首先保证，主持人的态度亲切、测试前随意聊聊彼此熟悉、测试提纲清晰全面。另外，以下几点也对保证测试的信度和效度很重要。

1. 不要忽略异想天开

脑暴中要求彼此不能批评，在进行访谈或测试中，也不能对用户某些操作做出评论，否则用户很有可能隐藏内心真实的感受。关注并记录用户出错，但是用户出错时态度要中立。通常，用户在体验的真实的原型后，会产生很多看似异想天开的诉求，有些虽然在当前不能实现，但是会为未来发展提供很多思路 and 方向。所以，我们要积极鼓励用户进行思维发散。

2. 前后验证、竞品比对

在测试完成后，可以加上一个总体调查问卷，一者让用户对自己体验的各个功能点有一个回顾和比较，同样也可以验证用户体验过程的态度和最终的态度是否具有 consistency。如果存在不一致，应该进一步追问理由，确定用户的真实想法。

测试时，让用户体验竞品，并作出比较，也是发现有效信息的途径。

3. 敏锐观察

测试中，除了按照已定的提纲进行问答之外，过程中还要敏锐的观察用户一些细微的表情、停留、思考。不但要了解用户对个功能点如何评价的，还要知道用户做某一任务过程中，是怎么思考、计划、实施的，用户的第一反应、习惯性的操作、思维路线的作用远远大于单纯的评价。用户任务完成之后，要追问用户如此操作的原因。

4. 记录原话并习惯性确认

测试结论要有用户的原话支持，不能轻易的改变用户的表述。和用户交流过程中，要习惯性的问：“请问你的意思是……？”“我这样理解你的意思，你看对么……？”以保证测试结论的效度。

5. 必要时进行入户调查

首先，入户调查会大大减少外界环境的影响，用户在自己的空间中，会更真实的反映常见的问题。其次，入户调查一般是在用户画像提取出来之后，按照 用户画像描述的属性，有意识有针对性去挑选具有某些典型属性的对象进行深入、全面、系统调查（典型调查），比如某一产品的目标用户，他们反映的问题，代表 性强，往往有以一当十的功效，避免了非目标用户信息造成的干扰。

6. 用户条件与数量

参与测试用户根据目标用户特征选择。

一般衡量测试是否需要继续进行的方法是：看是否发现新的问题，如果有新的问题，就应该继续，反之，可以结束。

Neilson 研究表明，5 名用户的测试可以发现 85% 的可用性问题。而在我们在以往的可用性测试经验中，用户数一般定为 6 个，基本上能发现全部问题。当然任何数字都只是一个参考，用户数量最好根据具体的测试情况（衡量时间、资源、投入产出比）而定。总之，关键在于是否有新的问题出现。

信度效度贯穿整个用户研究过程，肯定会有很多没有考虑到的地方，还请各位轻轻拍砖。

参考

1. 朱远程、马栋，“谈结构方程模型的应用策略”，[企业管理]，2010
2. 李乐山教授 2010 腾讯演讲
3. <http://www.useit.com/>
4. 刘金兰等译，“美国顾客满意度指数”[管理学报]，2005