实验结果汇总

符号：

Etime：Estimate Time

Dtime: Dwell Time

Sat: Satisfaction

Midsat: Medium Satisfaction

Unsat: Unsatisfaction

实验 1.1 研究 temporal relevance，satisfaction 单个变量对于感知时间的影响。

基本假设：在 high temporal relevance 条件下的用户，比 low temporal relevance 条件下的用户倾向于估计更长的时间。

实验：我们在 satisfaction 设置相同的情况下，比较同一个 task 上 high temporal relevance 和 low temporal relevance 条件下的差值 Avg(Etime-Dtime) 以及差值与 Avg((Etime-Dtime)/Dtime)。

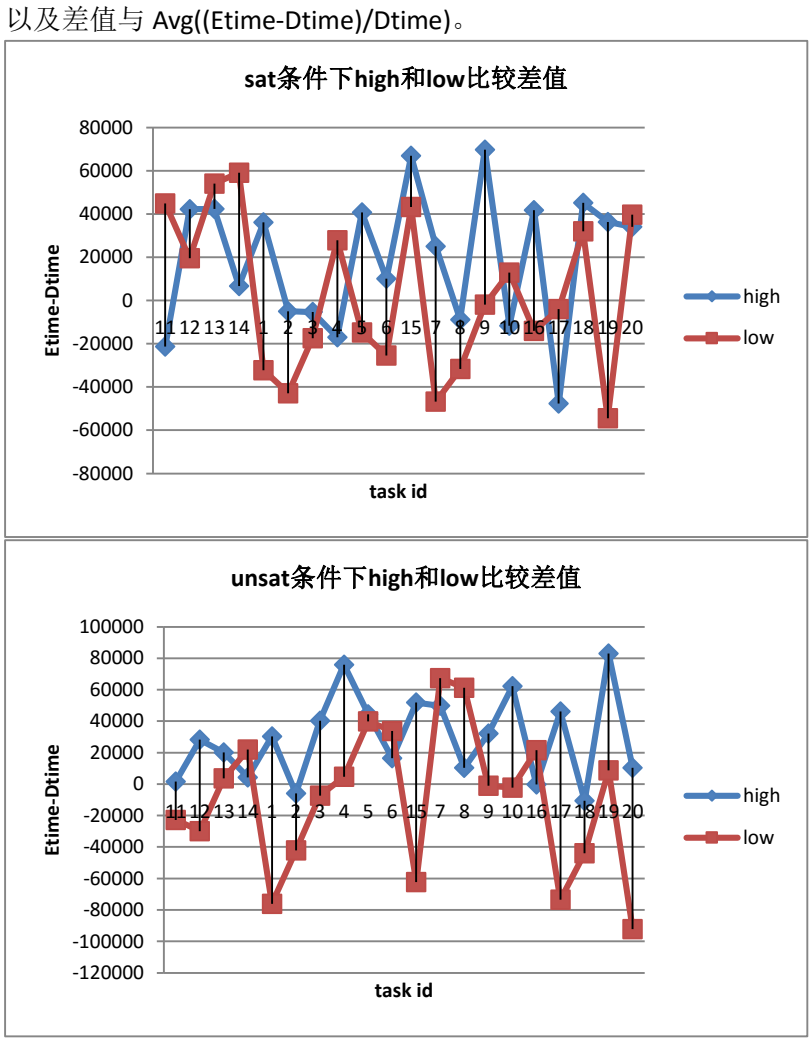


图 1 在不同满意度条件下，High/Low Temporal Relevance 条件下用户估计时间的情况

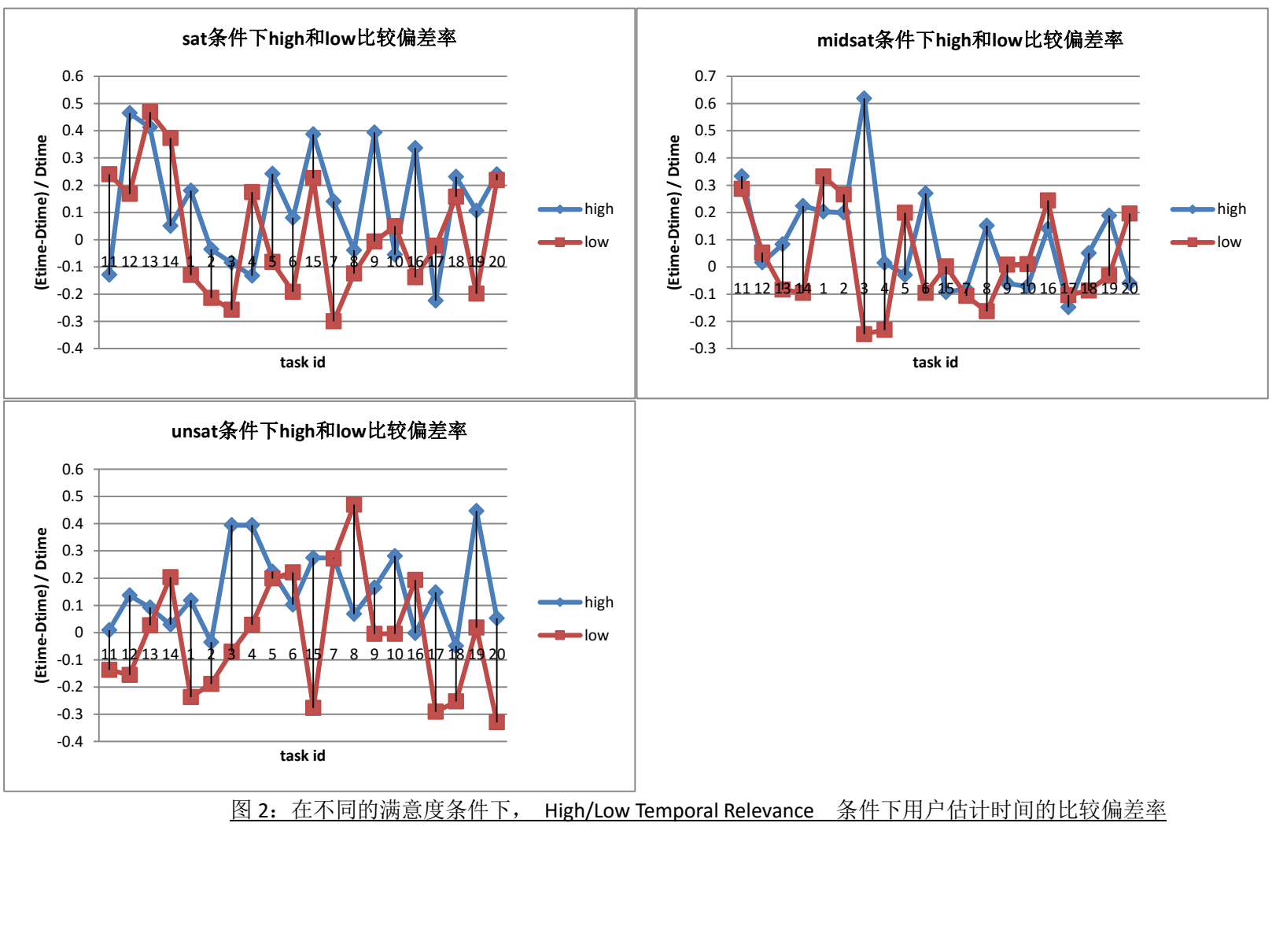


图 2：在不同的满意度条件下， High/Low Temporal Relevance 条件下用户估计时间的比较偏差率

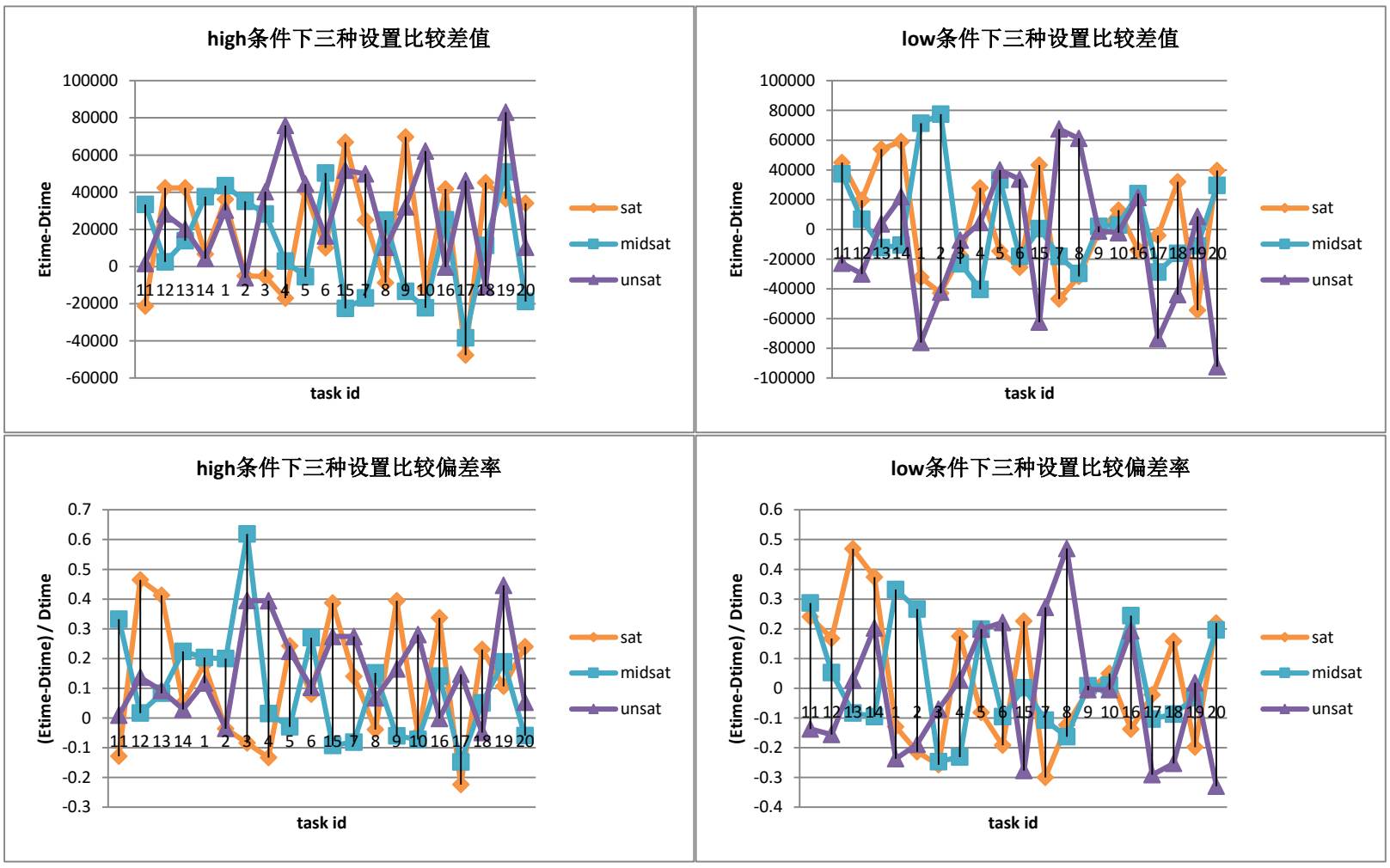
图中，横轴是 task 的序号，每一个点上是 5 个用户的用户的平均值。其中，task11—14 是 time critical 的 query。直观看来，

1. High temporal Relevance 条件下，用户倾向于估计的时间更长，但是这个情况在不同的满意度设置下，在不同的 task 上并不一致。图 1 中
   在 Sat 条件下，11,13,14,4,10,17,20 与假设不一致；
   在 Midsat 条件下,11,13,1,2,5，15,9,10,20 与假设不一致；
   在 Unsat 条件下，14,6,7,8,16 与假设不一致；
2. 从差值和偏差率来看，high 条件下的值大体上比 low 条件下的值大，这从一定程度上说明了 high 条件下用户在 Dtime 的基础上倾向于估计更长的时间。不过这个现象并不十分明显，可能的原因是没有考虑到 user 的因素，这里的数据都是所有满足条件的 user 的平均值，但每个 user 并不是把所有的 task 都做过一遍。
3. 在图 1 和图 2 所示的实验中，暂时看不清 Sat 控制的效果。我们讨论到一种情况，可能会导致 sat 的控制失效，比如说有一个结果，在第 7、第 8 位可以找到一个 relevance 的结果，但是 reverse serp 之后反而到了前面。
4. 上面的实验结果事实上认为用户是同质的，即不同的用户对时间估计的偏好是差不多的，这个肯定是有问题的，在后面有一个相关的实验。

实验 1.2 研究在不同的 Temporal Relevance 设置下，不同的 Satisfaction 条件对用户估计时间的影响。

基本假设：用户在越不满意的情况下倾向于估计更长的时间。

实验：我们在 temporal relevance 设置相同的情况下，比较同一个 task 上 sat，midsat 和 unsat 条件下 Etime 和 Dtime 的差值，以及差值与 Dtime 的比值（偏差率）。



从差值和偏差率来看，sat，midsat 和 unsat 三种条件下看不出明显的大小关系，我们的假设并不成立。可能的原因有两方面，

第一，我们对于 sat，midsat 和 unsat 的设置失效。实验中我们是颠倒搜索结果的排序来进行设置，但由于我们的系统本身对搜狗的搜索结果进行了过滤，而且有一些 task 在最靠前的位置反而并不能找到好的结果，这使得 sat，midsat 和 unsat 的标准并不可靠；

第二，在实验这种环境下，用户对于 task 的满意度的理解和日常的搜索环境可能是不一样的，有可能用户花了很大功夫找到了一个很满意的答案，他就会觉得很满意，我们给用户设定的信息需求不一定能反映他的真实心理，他更多只是把这当成一个任务。

总结：这一部分主要研究 temporal relevance，satisfaction 单个变量对于时间感知的影响。就 satisfaction 来说，并没有找到任何有价值的证据来验证我们的假设，而 temporal relevance 方面，实验数据一定程度上说明了 high 条件下用户在 Dtime 的基础上倾向于估计更长的时间，不过这个现象目前还并不十分明显，可能的原因是没有考虑到 user 的个性化因素。

实验 2: 研究 task 本身是否具有一定的属性？比如会倾向于让用户估计时间更长或者更短。

实验：在整个实验中，每个 task 会被做 30 次，其中 high temporal relevance 和 low temporal relevance 条件下各会被做 15 次，我们比较两种条件下 task 的 Etime 比 Dtime 长的比例。

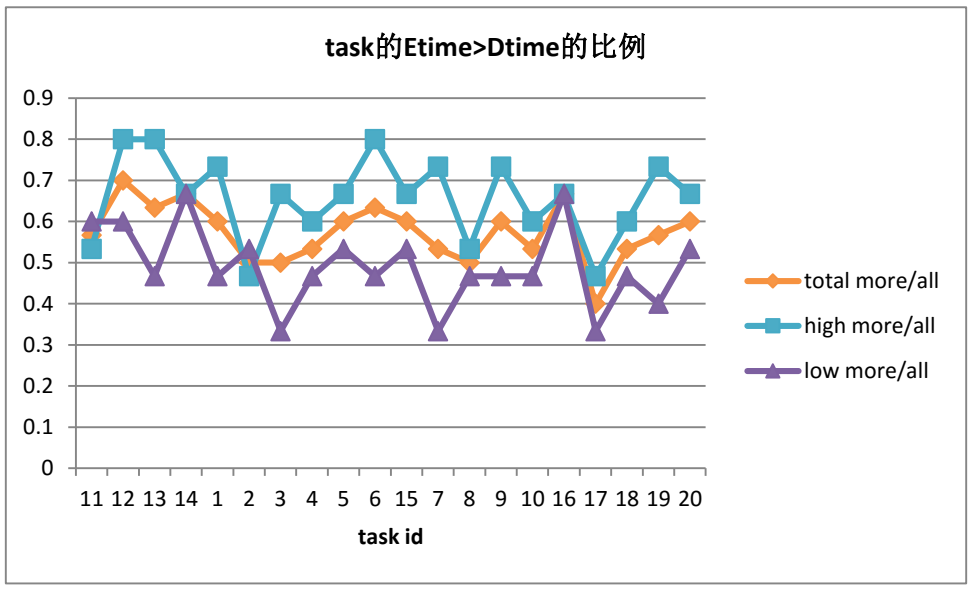


图 3. 在不同的设置下，Etime > Dtime 的比例

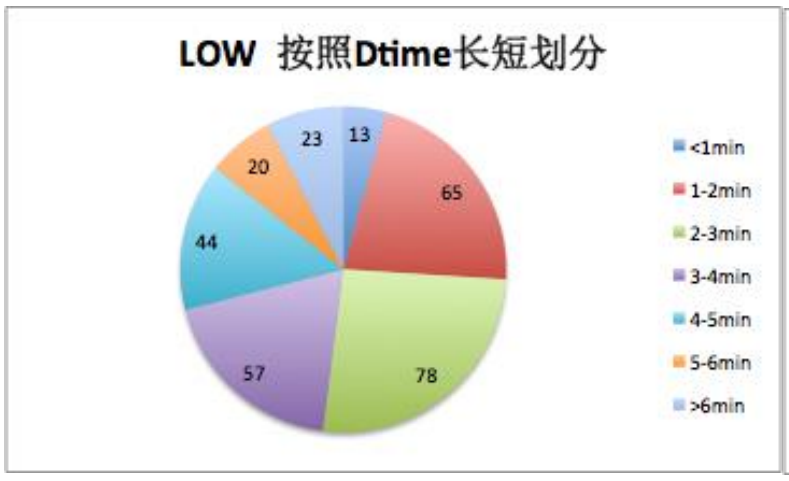
可以看出，对于同一个 task 来说，high temporal relevance 条件下，Etime>Dtime 的用户（实验样本）比例更高，这说明 high temporal relevance 条件下用户更倾向于把时间估计得比真实时间更长。

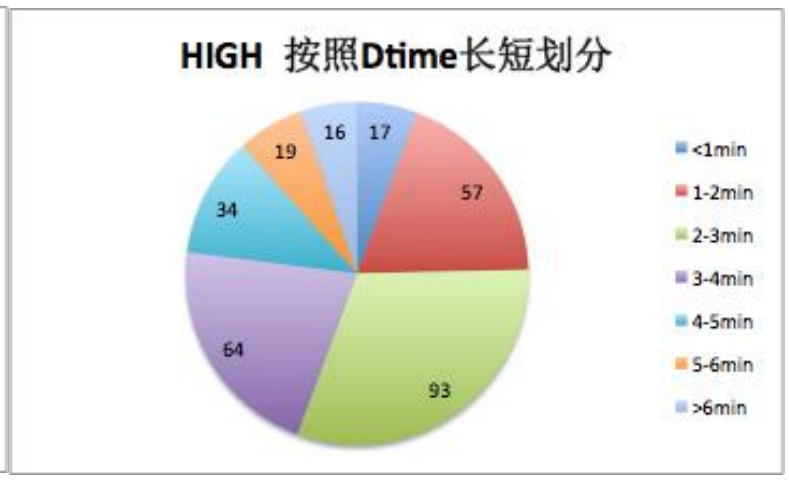
实验 3：Dtime 对 Etime 的影响。

基本假设：dwell time 的长短会对 etime 的估计产生影响，比如 dwell time 短的时候，影响比较大。

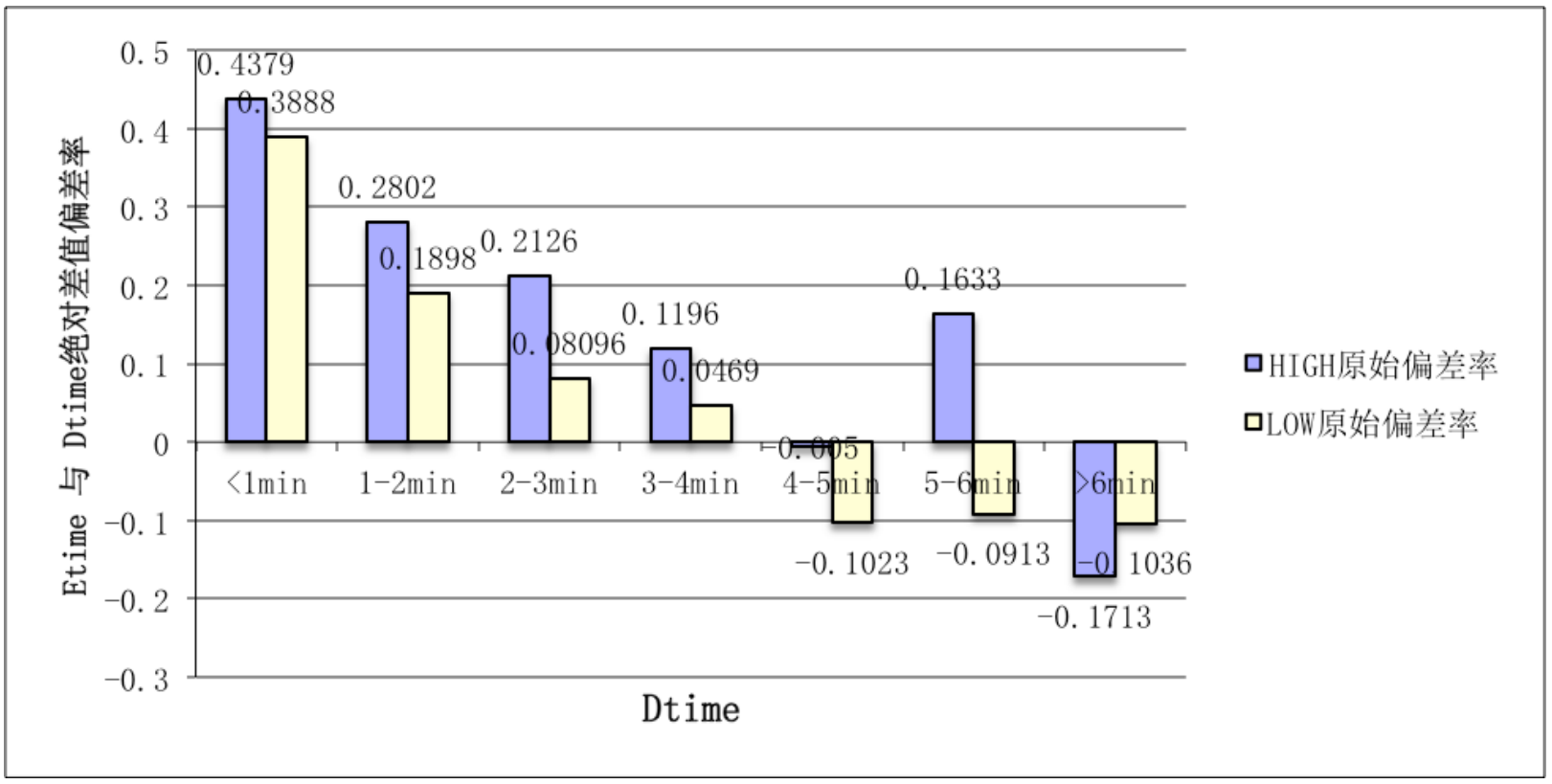
实验：我们分别在 high temporal relevance（后简称 HIGH）和 low temporal relevance（后简称 LOW）下，分析 dwell time（Dtime）和估计时间偏差率的关系，看看在这两种情况下 Dtime 是不是都会对偏差率产生影响。如果有影响的话，会产生什么样的影响。

1. 首先，根据 Dtime 的长短，每间隔一分钟划分一个区间，共划分出了 7 个区间。我们先统计 HIGH、LOW 各 300 条结果的 Dtime 在各个区间的分布情况，如下两图所示。很容易发现，用户完成实验的 Dtime 主要集中在 1-5min 的区间上（均>80%）。



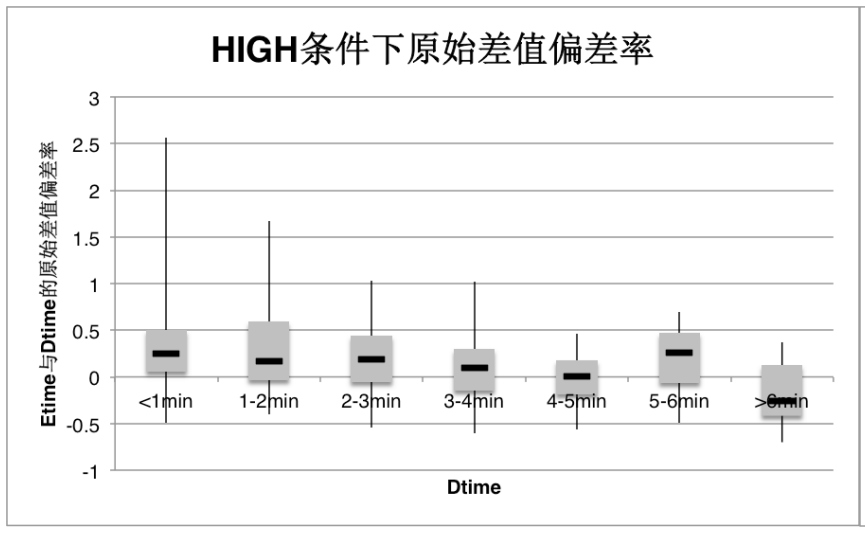


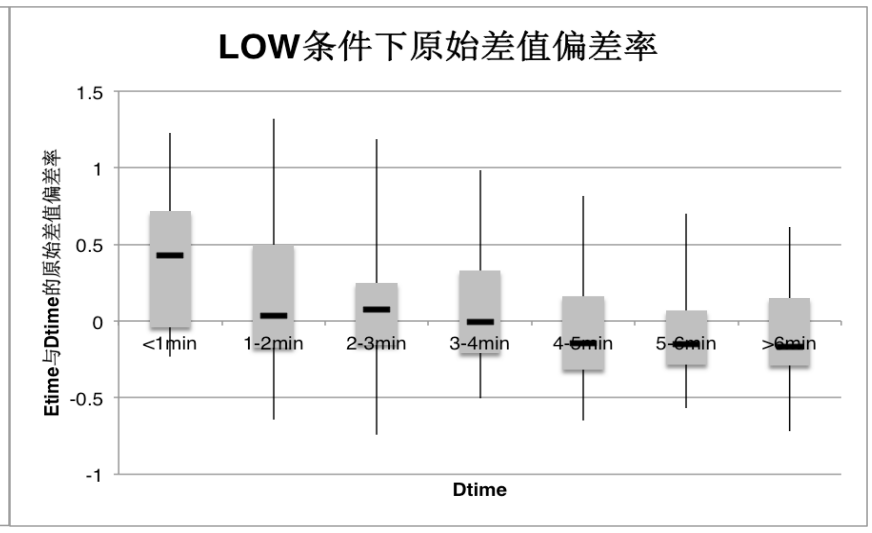
2. 在根据 Dtime 划分出的每个小区间中，计算 Etime 与 Dtime 二者差值偏差率的平均值，即（Etime - Dtime）/ Dtime 的平均值。这不仅可以让我们看到估计的偏差率，还可以看到用户是更容易将时间估计长还是估计短。比较不同区间之间，以及 HIGH、LOW 两种设置下的偏差率。
绘制出的图像如下图所示。只考虑 Dtime 分布集中的 1-5min 区间，可以明显看出随着 Dtime 增长，偏差率逐渐减小，即时间越长估计越准确。而且可以看出随着 Dtime 增长，用户倾向于将时间估计短。
我们还观察到，在每个区间内，都是 LOW 比 HIGH 偏差率小，即估计得更加准确一些。原因应该是 HIGH 条件下用户更容易估计出更长的时间。



3. 由于平均值并不能反映出每个区间的整体情况，所以接下来采用箱体图来进行进一步的数据展示。
分别看 HIGH 条件下和 LOW 条件下的偏差率，在 1-5min 区间基本可以看出箱体的整体下移趋势。也就是说上面得出的随着 Dtime 增加，用户

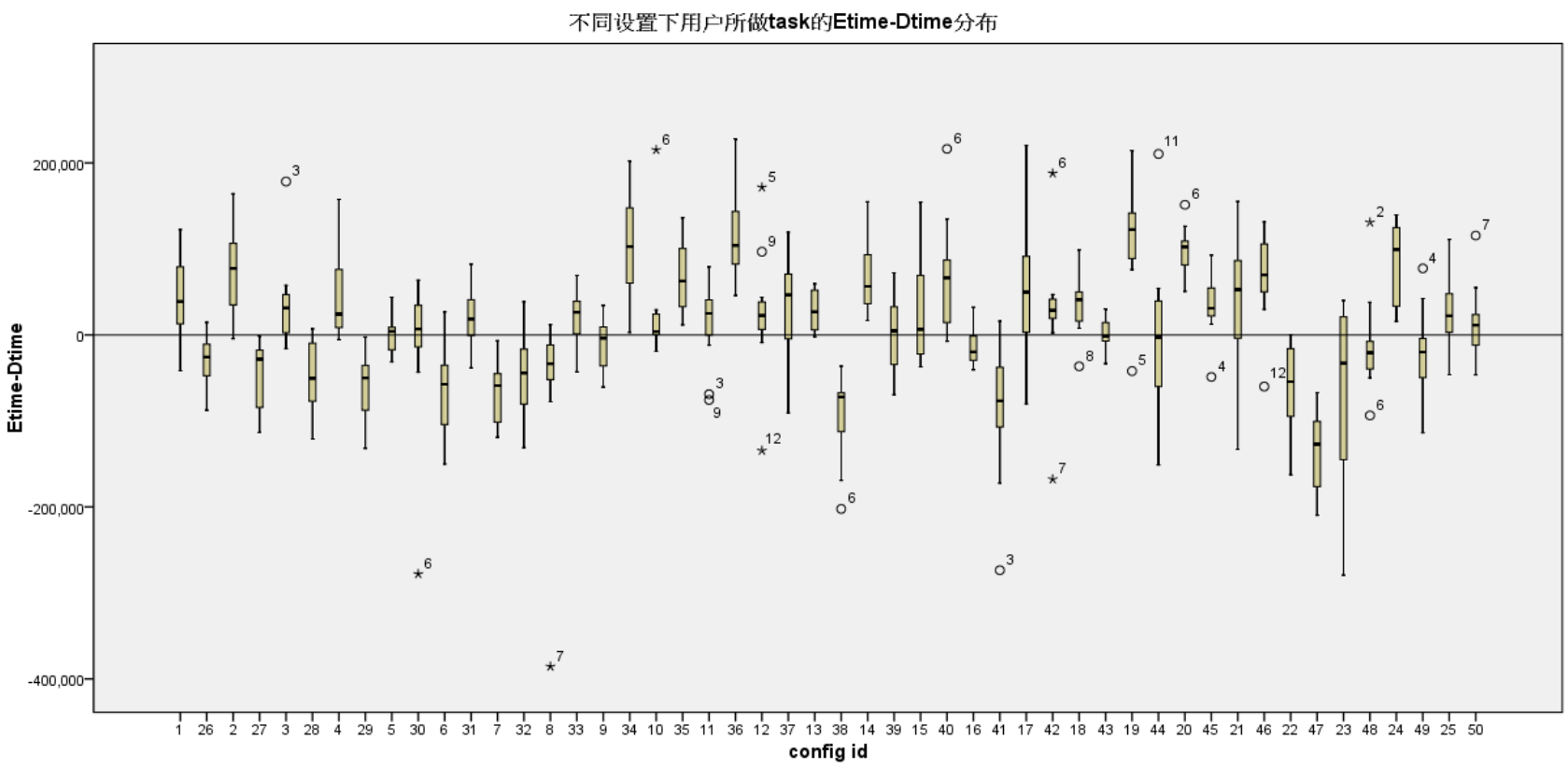
的时间感知越来越准确，这一结论是比较可靠的。
上面得出的第二个结论，LOW 比 HIGH 估计得更加准确，这一点只能在 1-4min 区间中通过中位数看出来，并不十分显著。
还可以明显看出，LOW 在每个区间的数据分布范围都比 HIGH 的要大很多，这比较容易理解，因为 HIGH 条件有一定的时间参考，个人估计时间的差异性会被削弱一些。

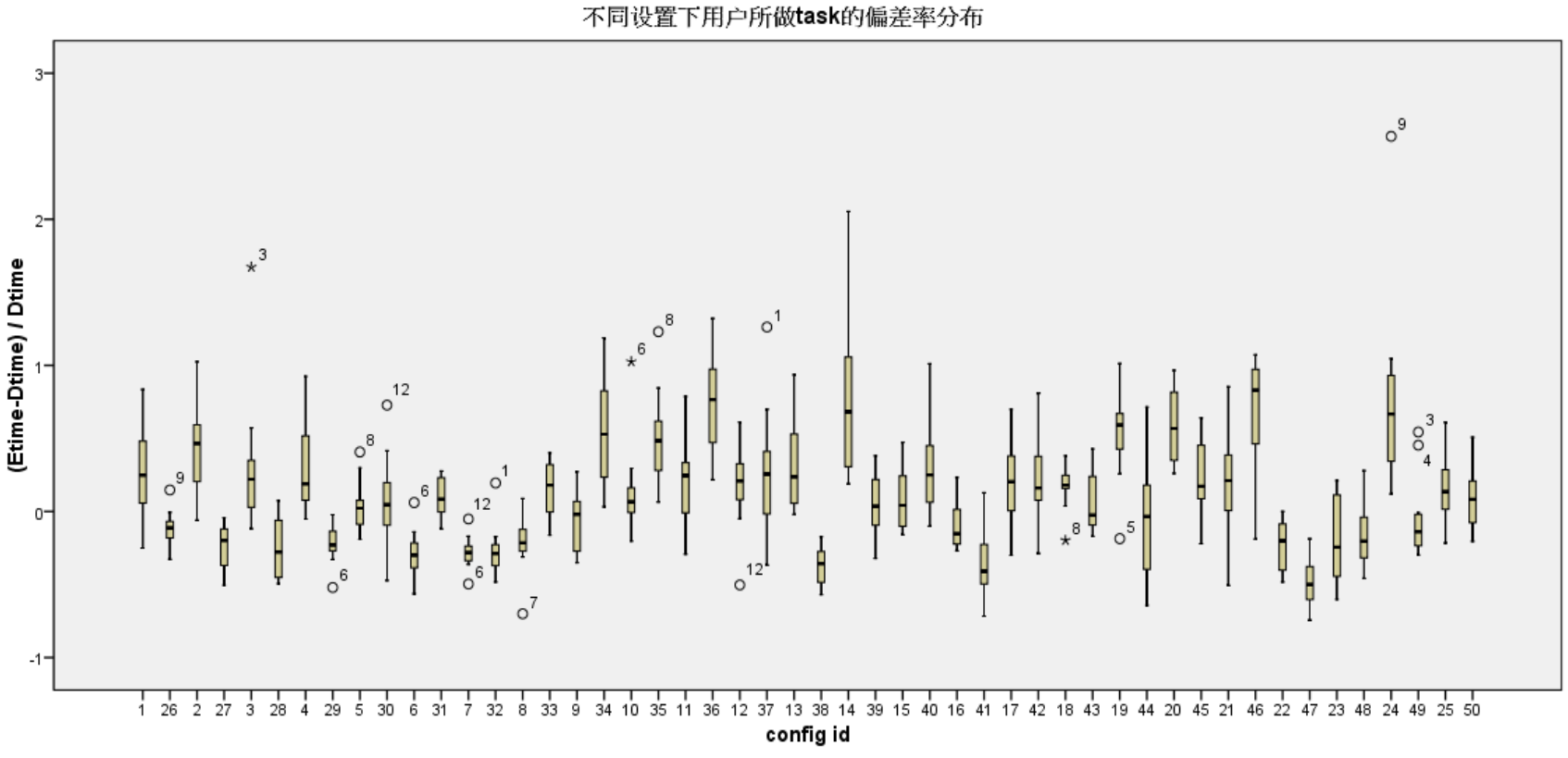




总结：通过实验我们验证了提出的假设，不同的 dwell time 确实会实验者估计时间的偏差产生影响。用户 Dtime 集中分布于 1-5min，在这个区间内，用户完成任务的实际时间越长，对时间的感知就更准确。并且，HIGH 条件会有效抑制用户本身估计时间的个人差异性，数据分布范围更小。

一、研究不同 temporal relevance 条件下不同 user 的时间感知的情况
假设：high temporal relevance 条件下的 user 比 low temporal relevance 条件下的 user 倾向于估计更长的时间
实验：我们对其他设置相同，只有 temporal relevance 设置不同的 user（设置 1 和 26, 2 和 27, …, 25 和 50）所做的 12 个 task 的 Etime 和 Dtime 的差值以及差值与 Dtime 的比值（偏差率）进行对比，看二者是否有明显区别，为了直观看出 user 所做的 12 个 task 的整体差异，用箱线图表示。





图中横轴相邻的两个设置（如 1 和 26）只有 temporal relevance 设置不同，可以看出基本上两个 user 估计的 12 个 task 的差异是比较明显的，但我们的假设并没有得到验证，因为虽然差异明显，但有的是 high 条件下差值和偏差率大，有的是 low 条件下差值和偏差率大，这不能直接说明 temporal relevance 这个因素对感知时间的影响，这可能跟不同的设置有关，即可能在有些设置下 high>low，在有些设置下 low>high，这说明设置本身也会对时间感知带来影响，要预测时间感知需要考虑到设置方面存在的变量。不过值得注意的一点是 user 与 user 之间估计行为的差异是比较明显的，这说明我们考虑 user 个性化的因素是很有必要的。