Stationary and convergence properties of 'Up-and-Down' methods

Assaf P. Oron

Department of Statistics, University of Washington, Seattle

assaf@u.washington.edu

September 27, 2006

Abstract

We examine three modifications of the 'up-and-down' (U&D) median-finding method for binary-response experiments. These modifications - 'biased-coin design' (BCD), 'k-in-a-row' (KR) and 'group up-and-down' (GU&D) - target non-median threshold percentiles. Their stationary and convergence properties are compared theoretically and numerically. KR is found to be superior to BCD in all respects. KR converges faster than GU&D, but the latter has smaller stationary bias. Numerical convergence calculations indicate that sample sizes of 10 or less used in some fields for median estimation are overly optimistic, as are sample sizes of < 30 for finding the 30th or 20th percentiles using any of the three above-mentioned methods. The '3+3' method, commonly used for Phase I clinical trials, appears to converge even more slowly. We also conclude that using only $4-6$ treatment levels is sub-optimal, due to a convergence-precision trade-off. Instead, $8-10$ levels are recommended.

Keywords: Adaptive Staircase; Phase I Clinical Trials; Quantile Estimation; Random Walk experimental design; Sensory Threshold; Up-and-Down;

1 Introduction

In many binary-response experiments, researchers are not interested in finding the entire treatment-response curve, but only a single representative value, which can be described via a percentile of the underlying threshold distribution. This percentile can represent (for example) the sensory threshold, the LD_{50} of a toxic substance, or the maximum-tolerated-dose (MTD) of a medication. The 'up-and-down' (U&D) sequential method, existing since the 1940's (Dixon and Mood, 1948; von Bekesy, 1947; Anderson et al., 1946), was designed to estimate the median threshold, under constraints of moderate sample size and a discrete set of treatment levels. It is still in use today in its original application - finding the median sensitivity of explosives to shock loading (Chao and Fuh, 2003). It can also be found in a wide array of applied research fields - including fatigue testing in metallurgy and material science (Lagoda and Sonsino, 2004), testing of dental restorative materials (Scherrer et al., 2003), breakdown voltage estimation in electrical engineering (Komori and Hirose, 2000), animal toxicity response (Sunderam et al., 2004), and anesthesiology (Capogna et al., 2001; Drover et al., 2004). In several applications, U&D is considered a standard method (ASTM, 1991; OECD, 1998; NIEHS, 2001). However, U&D is still not as thoroughly studied as other standard statistical methods.

Methodological research during U&D's early years (Brownlee et al., 1953; Derman, 1957; Wetherill, 1963; Dixon, 1965; Wetherill et al., 1966; Tsutakawa, 1967) generated accepted practices for design and estimation. After a two-decade lull, interest in U&D has been gradually increasing since the late 1980's - mostly in the context of Phase I clinical trials (Storer, 1989). This recent research has yielded results that incorporate advances in the study of Markov chains (Durham and Flournoy, 1994, 1995; Durham et al., 1995; Gezmu, 1996; Giovagnoli and Pintacuda, 1998; Ivanova et al., 2003; Bortot and Giovagnoli,

2005; Gezmu and Flournoy, 2006) and in estimation (Komori and Hirose, 2000; Chao and Fuh, 2001; Stylianou and Flournoy, 2002; Stylianou et al., 2003). Much of this progress is summarized by Ivanova and Flournoy (2006). Unfortunately, quite often recent recommendations are not properly compared with the older set of standards, which are still used by researchers in other applications. Meanwhile, some of the method's basic properties are seldom discussed - most notably convergence rates. This leads to increasing discrepancies in U&D design and estimation across different fields.

Sample size is a case in point. Dixon and Mood (1948) originally recommended a sample size of $N \approx 40$ for median estimation. Shortly afterwards, there was a trend to tailor the method for smaller and smaller sample sizes, $N < 10$ (Brownlee et al., 1953; Dixon, 1965). Nowadays, some military industries still use $N \approx 40$ for explosive testing (H. Dror, personal communication). Some groups in anesthesiology use a fixed sample size of 30 for median estimation (Capogna et al., 2001; Camorcia et al., 2004), while many researchers in the same field and elsewhere use $N < 10$ for the same goal (Lichtman, 1998; Sunderam et al., 2004; Drover et al., 2004). These differences are primarily a result of varying local traditions, rather than different needs or theoretical insight.

The work reported here has been motivated by an anesthesiology experiment to find the 20th percentile of pain response to a mixture of anesthetic agents (Oron, 2004). Modifications of U&D targeting non-median percentiles have been proposed since the 1950's. This article focuses on three such methods: 'biased coin design' (BCD) (Durham and Flournoy, 1994, 1995; Durham et al., 1995, 1997),'k-in-a-row' (KR) (Wetherill et al., 1966; Gezmu, 1996; Ivanova et al., 2003), and 'group up-and-down' (GU&D) (Anderson et al., 1946; Tsutakawa, 1967; Gezmu and Flournoy, 2006). These methods - arguably the simplest non-median U&D modifications - are recently being considered for Phase I clinical

trials, especially when researchers are not inclined to make parametric assumptions (Storer, 1989; Rosenberger and Haines, 2002; Ivanova and Flournoy, 2006). Interestingly, the least-studied of the three (KR) is the only one having a substantial track record in practice: it has been widely used to estimate sensory thresholds since the 1960's (Wetherill and Levitt, 1966; Treutwein, 1995; Garcìa-Perez, 1998; Marvit et al., 2003).

The same discrepancies in usage and theory that plague median-targeting U&D can be observed with these non-median methods. In vision research where KR is used, actual sample sizes range anywhere from ~ 20 to > 100 for the same task (Garcìa-Perez, 1998). Methodological clinical-trial oriented studies focus on sample sizes of $15 - 35$, influenced not so much by statistical theory as by that specific application's practical and ethical constraints (Durham et al., 1995; Ivanova et al., 2003). Numerical evidence from two concurrent studies (Bortot and Giovagnoli, 2005; Gezmu and Flournoy, 2006) and from my own preliminary work (Oron, 2005), indicate that this latter range is overly optimistic. Similar problems exist with regards to estimation (see e.g. the different approaches recommended by Garcìa-Perez (1998) and by Stylianou and Flournoy (2002)).

Perhaps before everything else, the simple question of which U&D modification to choose for a given application, has not been properly addressed. Since the 1950's researchers have creatively proposed many dozens of non-median U&D modification (e.g., Wetherill et al. (1966); Storer (1989); Garcìa-Perez (1998); Ivanova et al. (2003); Bortot and Giovagnoli (2005)). Yet, different U&D methods are compared almost exclusively via limited-scope simulations, and therefore the question "which U&D method is better?" is still unanswered.[1] There are also many innovative non-U&D percentile-finding designs (O'Quigley

[1]The only exception that comes to mind is Bortot and Giovagnoli (2005), who prove that BCD is optimal among first-order Markovian methods. But applied researchers do not care about a method's Markov-chain order; they would like to know which method is best for estimating a given percentile, under certain practical constraints.

et al., 1990; Rosenberger and Grill, 1997), and one may legitimately question whether U&D is even the right platform for this task. But a comparison of U&D with other approaches is essentially meaningless, if the U&D design and estimation methods used can be easily outperformed by other available U&D methods, of which the researchers were not aware.

The comparison presented in this article aims to provide a more comprehensive benchmark and framework for this issue. Additionally, the comparison is used in order to fill some gaps in U&D's theoretical foundation, starting from the basics - stationary and convergence properties. The emphasis has been on obtaining proofs and theoretical analysis whenever possible, and complementing them with numerical results where needed. The original median-targeting 'simple U&D' (SU&D) serves as a starting point in each section. Section 2 describes the methods. Section 3 compares the methods' stationary distributions, and Section 4 discusses and compares convergence rates. Final discussion and conclusions appear on Section 5.

2 The Methods

2.1 Terminology and Notation

Let X be a sequentially-determined experimental treatment with a binary response $Z(X)$, which has a response threshold CDF $F(X)$. We assume F is continuous and strictly increasing. Researchers look for $F^{-1}(p)$, the $100p$-th percentile of F. Treatments are administered sequentially: the set of actual treatments and responses is indexed $x_i, z_i,\ i = 1 \dots n$. Following trial i, the next treatment X_{i+1} is determined by some subset of all previous treatments and responses $\{x_1 \dots x_i, z_1 \dots z_i\}$, and by the specific method's transition rules. In U&D methods, experimental treatments are restricted to a fixed set of levels

$l_m, m = 1, \ldots$. Here we also assume (for convenience of analysis) that the set $\{l_m\}$ is finite so that there is a maximum boundary level l_M, and that levels are uniformly-spaced (or log-spaced), with level spacing s. The (hypothetical) level corresponding to the target percentile is denoted l_{m^*}, such that $l_{m^*} = F_p^{-1}$. [2] The following discussion focuses on targets below the median, but all results can be trivially translated to targets above the median.

This terminology allows us to generically define an U&D design within the context of this paper. The following definition is in the spirit of Potter (2002), Ivanova and Flournoy (2006) and others.

Definition 2.1 *(i) An* **'Up-and-Down' design** *is any discrete-outcome sequential experimental design with treatments and responses $\{x_n\}, \{z_n\}$, respectively; with fixed, discrete treatment levels $\{l_m\}$ and sequential transition rules limited to transitions of one level up, one level down, or remaining at the current level.*

(ii) In a **rule-based or nonparametric U&D design***, the transition rules are functions of $\{x_n\}, \{z_n\}$ and possibly also ϕ, a set of fixed parameters.*

(iii) In a **model-based or parametric U&D design***, the transition rules are functions of $\{x_n\}, \{z_n\}, \phi$, and a statistical model $f(\{x_n\}, \{z_n\} \mid \theta)$, where θ is a set of data-estimable parameters.*

An example for a model-based U&D is the restricted CRM suggested by Goodman et al. (1995). The methods examined here all belong to the former, rule-based group. Therefore from here on, any reference to U&D implies an rule-based U&D, unless otherwise specified.

[2] Since F is assumed to be continuous strictly monotone, with probability 1 l_{m^*} is not part of the set of possible treatments, i.e. m^* is non-integer. However, the notion of m^* will be useful later on.

2.2 Transition Probabilities, Balance Equations and Stationary Formulae

The original median-targeting SU&D starts at an arbitrary level, and then moves up or down 1 level, following *'no'* or *'yes'* responses, respectively. This makes SU&D a lattice random walk, with 'up' and 'down' transition probabilities $p_m \equiv Pr\left(X_{i+1} = l_{m+1} \mid X_i = l_m\right) = 1 - F(l_m)$ and $q_m \equiv Pr\left(X_{i+1} = l_{m-1} \mid X_i = l_m\right) = F(l_m)$, respectively.[3] Assuming $F(l_m) \in (0,1)\ \forall m$, there is a stationary distribution (Tsutakawa, 1967; Durham and Flournoy, 1994):

$$\begin{aligned} \pi_m p_m &= \pi_{m+1} q_{m+1},\ m = 1 \ldots M-1 \\ \gamma_m \equiv \frac{\pi_{m+1}}{\pi_m} &= \frac{p_m}{q_{m+1}} = \frac{1 - F(l_m)}{F(l_{m+1})} = \frac{1 - F(x)}{F(x+s)}, \end{aligned} \tag{2.1}$$

with straightforward normalization to determine π_1. γ_m, hereafter termed 'the stationary distribution profile', is monotone decreasing in m. Therefore, the stationary distribution has a single mode, including the possibility that the mode is on a boundary (Durham and Flournoy, 1994). The U&D treatment sequence is known as a random walk with a central tendency (see e.g., Hughes (1995), Ch. 3.4). Durham and Flournoy (1994) prove that the stationary mode is at most one spacing interval away from the median.

Derman (1957) suggested modifying U&D via randomization in order to target non-median percentiles. However, the method now known as **'biased-coin' design (BCD)** was developed independently by Durham and Flournoy (1994, 1995). It is similar to SU&D, except that after a *'no'* response the next treatment is determined using a random draw: with probability $\Gamma/(1-\Gamma)$ we go up one level, otherwise the level is unchanged ($\Gamma \in (0, 0.5]$). Now r_m, the

[3] On the boundaries, trivial 'reflecting' conditions, i.e. $q_1 = 0, p_M = 0$, are imposed; this type of boundary conditions is assumed throughout the article, but explicit boundary details are omitted for brevity.

probability of remaining at the same level for another trial, can be nonzero. The resulting transition probability rule as a function of F is

$$\begin{aligned} p_m &= [1 - F(l_m)]\,\tfrac{\Gamma}{1-\Gamma}, \quad && m = 1 \ldots M-1; \\ r_m &= [1 - F(l_m)]\,\tfrac{1-2\Gamma}{1-\Gamma}, \quad && m = 2 \ldots M-1; \\ q_m &= F(l_m), && m = 2 \ldots M \end{aligned} \quad . \tag{2.2}$$

The stationary distribution obeys (Durham and Flournoy, 1995)

$$\gamma_m = \frac{1 - F(l_m)}{F(l_{m+1})} \frac{\Gamma}{1-\Gamma}. \tag{2.3}$$

Again, the walk has a central tendency and a single mode; it was shown (Durham and Flournoy, 1994) that the stationary mode is at most one spacing interval away from $F^{-1}(\Gamma)$, which is the method's designated target.

Another non-median U&D method is known as 'forced-choice staircase' or **'k-in-a-row' (KR)** (Wetherill et al., 1966; Gezmu, 1996). Here there is no random draw; instead, we must observe exactly k consecutive *'no*'s at a given level before moving up (the 'down' rule remains as in SU&D). For $k = 1$, KR reduces to SU&D. Even though KR is the most widely used non-median U&D method and is arguably easier to administer than BCD (requiring no random draw), its Markov chain properties are more complex and have rarely been studied. KR can be described either as a k-th order random walk with two sets of transition rules, only one of which applies at each trial (Gezmu, 1996):

$$\begin{aligned} &x_i = l_m, \text{ but } \exists j, i-k < j < i \text{ s.t. } x_j \neq l_m: && \begin{cases} p_m = 0 & \\ r_m = 1 - F(l_m), & m = 2 \ldots M \\ q_m = F(l_m), & m = 2 \ldots M \end{cases} \\ &x_i = x_{i-1} \ldots = x_{i-k+1} = l_m: && \begin{cases} p_m = 1 - F(l_m), & m = 1 \ldots M-1 \\ r_m = 0, & m = 1 \ldots M-1 \\ q_m = F(l_m), & m = 2 \ldots M \end{cases} \end{aligned} \quad ,$$

or as a random walk whose treatment chain $\{x_i\}$ is paired with a chain of internal states $\{y_i\}$, each internal state taking one of k possible values ($0 \dots k-1$, *cf.* e.g. Weiss (1994), Ch.6). Under this internal-state formulation the transition rules become

$$\begin{cases} p_{m,y} = 0, & m = 1 \dots M, \quad y < k-1 \\ r_{m,y} = 1 - F(l_m), & m = 2 \dots M, \quad y < k-1 \\ q_{m,y} = F(l_m), & m = 2 \dots M, \quad y < k-1 \end{cases}$$
$$\begin{cases} p_{m,k-1} = 1 - F(l_m), & m = 1 \dots M-1 \\ r_{m,k-1} = 0, & m = 1 \dots M-1 \\ q_{m,k-1} = F(l_m), & m = 2 \dots M \end{cases} \quad . \tag{2.4}$$
$$\begin{cases} X_{i+1} = X_i \Rightarrow Y_{i+1} = Y_i + 1 \\ X_{i+1} \neq X_i \Rightarrow Y_{i+1} = 0 \end{cases}$$

Theorem 2.1 *(i) KR's stationary distribution profile* $\{\gamma_m\}$ *is given by*

$$\gamma_m = \frac{F(l_m)\left[1 - F(l_m)\right]^k}{F(l_{m+1})\left\{1 - \left[1 - F(l_m)\right]^k\right\}} \tag{2.5}$$

(ii) Let the stationary marginal 'up' probability of a KR design be the weighted sum $p_m|_\pi = \sum_y \pi_{m,y} p_{m,y}$*, where* $\pi_{m,y}$ *is the stationary frequency of the state* $(X = l_m, Y = k-1)$*. Then*

$$p_m|_\pi = \frac{F(l_m)[1 - F(l_m)]^k}{1 - [1 - F(l_m)]^k} \tag{2.6}$$

.

(iii) KR's stationary distribution has a single mode.

Proof (i) Taking the internal-state approach, the balance equations between

adjacent treatment levels may be written as

$$\pi_{m,k-1}[1 - F(l_m)] = \pi_{m+1}[F(l_{m+1})]. \tag{2.7}$$

Note that in this approach, π_m, the stationary frequency of each level l_m, is found by summing over internal states: $\pi_m \equiv \sum_y \pi_{m,y}$.

Now, transitions between internal states of the same treatment level are possible only for single upward increments, with probability $1 - F(l_m)$. To maintain balance at stationarity, the internal state frequencies must obey $\pi_{m,y+1} = [1 - F(l_m)]\pi_{m,y}$, $y = 0 \dots k-2$, i.e. a diminishing geometric sequence. This enables us to calculate relative internal-state frequencies, and specifically the upper state:

$$\frac{\pi_{m,k-1}}{\pi_m} = \frac{\pi_{m,0}[1 - F(l_m)]^{k-1}F(l_m)}{\pi_{m,0}\left\{1 - [1 - F(l_m)]^k\right\}}$$

plugging back into (2.7), we obtain (2.5).

(ii) Since under stationarity $\pi_m p_m = \pi_{m+1} q_{m+1}$, this result is immediate from (2.5).

(iii) Differentiating $p_m|_\pi$ w.r.t to F shows that it is monotone decreasing as F increases. Since F itself is monotone increasing in m, and since $q_m|_\pi = F(l_m)$ is monotone increasing in m, γ_m is monotone decreasing and therefore there is a single stationary mode (Durham and Flournoy, 1994). $\square$

A somewhat different proof of result (i) appears in Gezmu (1996)'s unpublished dissertation.

What is KR's target? This question leads us to the issue of U&D targets in general, which has been evading definition. The intuitive idea that the U&D chain is drawn towards the percentile where 'up' and 'down' movements balance each other, was clearly on the minds of the originating researchers. However, rigorous definitions of target are hard to come by, perhaps because treatment

levels are discrete. Recent attempts (e.g., Giovagnoli and Pintacuda (1998); Gezmu and Flournoy (2006)) either limit themselves to a small subset of U&D methods, or use common English and approximate terms. I have found the following definition useful, generic and rigorous:

Definition 2.2 *Consider a rule-based 'Up-and-Down' design with a stationary distribution π, and with (marginal) transition probabilities $p_m|_\pi$ monotone decreasing and $q_m|_\pi$ monotone increasing in x. The design's* **target** $F^{-1}(p) \equiv l_{m^*}$ *is defined as the (hypothetical) treatment level l_{m^*} such that*

$$p_{m^*}|_\pi = q_{m^*}|_\pi \tag{2.8}$$

In words, the target is that treatment from which stationary 'up' and 'down' probabilities would be equal, had a design level been placed exactly there. It is straightforward to see that this definition does recover SU&D's and BCD's targets as the median and $F^{-1}(\Gamma)$, respectively. For KR, we take the marginal 'up' probability from (2.6), and equate it with the 'down' probability $F(x)$, to obtain the equation for the target:

$$\begin{array}{c} [1 - F(F_p^{-1})]^k = 1 - [1 - F(F_p^{-1})]^k \\ p \equiv F(F_p^{-1}) = 1 - \left(\frac{1}{2}\right)^{1/k} \end{array} . \tag{2.9}$$

This result has been known since the method's inception (Wetherill et al., 1966).[4] Unlike BCD, KR can target only a discrete set of percentiles; for $k = 2, 3, 4$, these are approximately $F^{-1}(0.293), F^{-1}(0.206)$ and $F^{-1}(0.159)$, respectively.[5]

KR's stationary mode shares the same basic properties of SU&D and BCD:

[4] Paradoxically, Wetherill derived KR's target in an erroneous way: he calculated the target of a GU&D method (to be presented below) that happens to have the same target.

[5] In sensory studies the method is inverted: one *'no'* response triggers an 'up' move, while k consecutive *'yes'* responses are required for a 'down' move. Hence the targets are $F^{-1}(0.707), F^{-1}(0.794), F^{-1}(0.841)$, etc.

Corollary 2.2 *Given a KR design targetted on $F^{-1}(p)$ such that $l_1 \leq F^{-1}(p) \leq l_M$. Then the stationary mode is at most one spacing interval away from $F^{-1}(p)$.*

Proof This result was proven for BCD by Durham and Flournoy (1994) and for all first-order U&D methods by Giovagnoli and Pintacuda (1998), using the monotonicity of γ_m. Since on Theorem 2.1 we saw that γ_m is monotone decreasing, the same conditions sufficient for these two proofs now exist for KR. □

In **group 'up-and-down' (GU&D)** (Tsutakawa, 1967) instead of a single trial, at each stage a cohort of k simultaneous trials with the same treatment are performed on different subjects. Probabilistically, GU&D does not use a binary-outcome trial, but rather the binomial outcome $Z_G(x) \sim Bin(k, F(x))$, the number of *'yes'* responses observed in the cohort. GU&D transition rules stipulate a move up if $Z_G < b$, a move down if $Z_G \geq t$, and staying at the same level otherwise (obviously, $0 \leq b < t \leq k$). As k increases, the increasing number of possible combinations of b and t provides a large variety of targets. Generic GU&D stationary formulae and other key properties are given by Gezmu and Flournoy (2006). Here we focus mostly on the GU&D subset with $b = 0, t = 1$ (hereafter: GU&D$_{(0,1,k)}$), whose transition probabilities are equal to those of SU&D performed on a transformed CDF $G(x) \equiv 1 - [1 - F(x)]^k$ (and with each single 'transformed trial' representing a group of k actual trials). GU&D$_{(0,1,k)}$ targets are identical to those of KR with the same k, thus enabling direct comparison between the three U&D 'flavors'. Additionally, as Storer (1989) and Ivanova and Flournoy (2006) note, the commonly used phase I cancer trial **'3+3' method** can be modeled as an initial GU&D$_{(0,2,3)}$ stage, changing into a more complicated design before stopping. Since convergence rate is governed mostly by the initial stage, the convergence of GU&D$_{(0,2,3)}$ will be examined in Section 4.

3 Stationary Distributions Compared

3.1 Peakedness

Due to limitations related to the discrete treatment set and to the unknown location of target relative to the fixed design levels, the notion of 'peakedness' is not easily converted to more familiar terminology used to describe dispersion (e.g. precision or variance), but it carries a similar meaning within the U&D context. The more 'peaked' the stationary U&D distribution is around its mode, the closer the method is to the ideal design, and percentile estimation precision will generally improve. Here is a rigorous definition of 'peakedness', following Giovagnoli and Pintacuda (1998).

Definition 3.1 *(i) For two 'Up-and-Down' designs,* **"all other things being equal"** *will mean that both target the same percentile, the threshold distribution F is the same, the treatment levels $\{l_m\}$ are the same and the initial conditions are the same.*

(ii) (Giovagnoli and Pintacuda, 1998) Let designs 1 *and* 2 *be two U&D designs. Then, all other things being equal, if $\gamma_m^{(1)} \geq \gamma_m^{(2)}$ for $l_m \leq F^{-1}(p)$ while $\gamma_m^{(1)} \leq \gamma_m^{(2)}$ for $l_m \geq F^{-1}(p)$ (i.e., design* 1*'s stationary distribution profile is steeper), then design* 1 *and its stationary distribution are called* **more 'peaked'**.

Thus, for example, to compare KR and BCD we need to examine

$$\frac{\gamma_m^{KR}}{\gamma_m^{BCD}} = \frac{F(l_m)\left(1-F(l_m)\right)^{k-1}}{1-\left(1-F(l_m)\right)^{k}} \frac{1-\Gamma}{\Gamma}. \tag{3.1}$$

Theorem 3.1 *For any $k > 1$, all other things being equal, KR designs are more 'peaked' than BCD designs.*

Proof On target, the ratio in (3.1) is exactly 1. Therefore, it suffices to show that (3.1) is monotone decreasing in F. Careful differentiation yields this re-

sult. □

Theorem 3.2 *For KR and GU&D$_{(0,1,k)}$ with the same k, neither design can be said to be more 'peaked' than the other.*

Proof The ratio analogous to 3.1 is

$$\frac{\gamma_m^{KR}}{\gamma_m^{GUD_{(0,1,k)}}} = \frac{F(l_m)\left[1-(1-F(l_{m+1}))^k\right]}{F(l_{m+1})\left[1-(1-F(l_m))^k\right]} \leq 1.$$

This inequality follows from the concavity of $1-(1-F))^k$ in F. GU&D$_{(0,1,k)}$'s stationary distribution profile is steeper to the left of target, while KR's is steeper to the right of target. □

Fig. 1 provides a numerical example of 'peakedness'. The differences between methods do not appear to be dramatic. On the other hand, if the spacing is too coarse (e.g. $M = 5$ on the left) the stationary peak is very shallow w.r.t. x, and the method's stationary properties offer only a modest improvement over standard, non-sequential treatment-response designs. As the spacing becomes finer (e.g. $M = 10$ on the right), a sharp peak forms: under stationarity, all 3 methods allocate $50-55\%$ of treatments to the two levels closest to target. Thus, finer spacing dramatically reduces the probability for treatments allocated too far from target. In the scenario illustrated on fig. 1, with $M = 10$ $Pr_\pi\left(x \geq F^{-1}(0.6)\right)$ (the 60th percentile is at $x \cong 0.7$) is 0.02 to 0.03 depending upon method, while with $M = 5$ the analogous probabilities are 0.12 to 0.14.

3.2 Bias of the Stationary Mean

The vast majority of U&D experiments are estimated using an empirical mean of (a subset) of $\{x_1 \ldots x_N\}$. In all these averaging methods, the are two implicit assumptions: that stationarity is a good approximation for the subset used, and

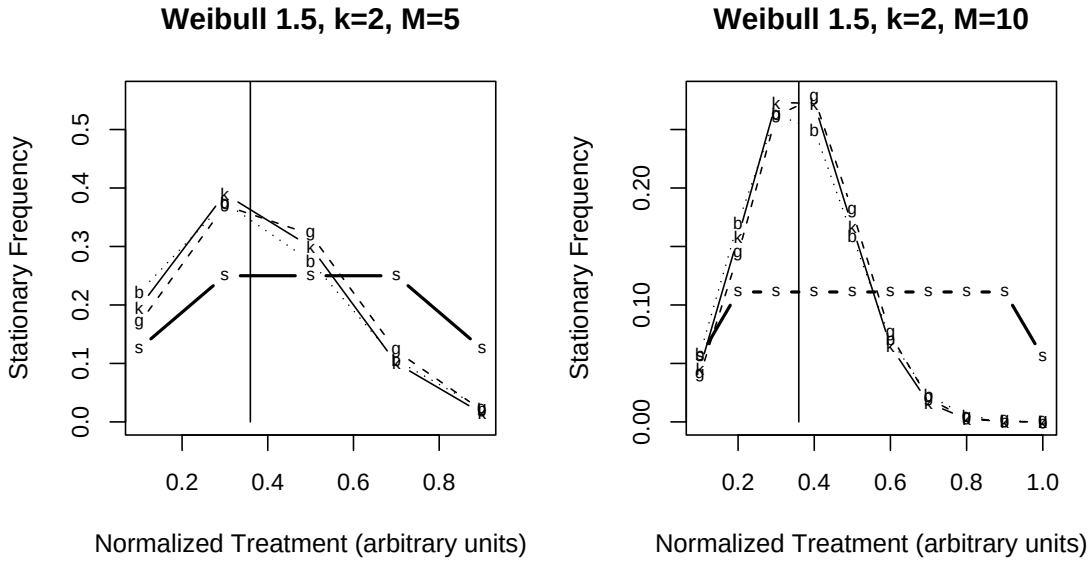

Figure 1: Stationary distributions for KR ('k' marks and solid lines), BCD ('b' marks and dotted lines) and GU&D ('g' marks and dashed lines) - all targeting $F^{-1}(0.293)$, marked with a vertical line. F is Weibull with shape parameter 1.5, and scale normalized so that $F(1) = 0.8$. Shown are a coarser design with $M = 5$ (left) and a finer design with $M = 10$ (right). Also plotted is a standard non-sequential treatment allocation of $1/(M-1)$ to interior levels and $1/2(M-1)$ to boundary levels ('s' marks and thick lines). The vertical axes are normalized to account for the factor 2 difference in M, so that the apparent area under the curves is similar in both plots.

that the stationary mean has negligible bias. In spite of the strong practical motivation to examine these assumptions, since the 1960's there has been no theoretical discussion of U&D's bias, and the issue was never examined from a Markov-chain perspective.

In the limit of infinitely fine design ($s \to 0$), the stationary distributions of BCD, KR and GU&D all tend to a delta function around target. With finite spacing, biases may be related to both s and the offset of treatment levels w.r.t. target. Dixon (1965) claimed without proof that the mean of SU&D treatments is an unbiased estimate of $F^{-1}(0.5)$ if F is symmetric. Indeed, if m^* is an integer or half-integer, the stationary distribution is symmetric too, and unbiasedness easily follows. In the case of a small offset in the design over a symmetric distribution, the mean is biased, but numerical calculations indicate that the bias is indeed negligible (data not shown).

We now examine SU&D bias for general threshold distributions. In the subsequent discussion, we assume for convenience that m^* is an integer (half-integer is equally convenient), and that F is twice continuously differentiable. One can then express the stationary mean as a telescopic series around target:

$$\begin{array}{l} \bar{l_\pi} \equiv \sum_{m=1}^{M} \pi_m l_m \cong l_{m^*} + s \sum_{j=1}^{U} j(\pi_{m^*+j} - \pi_{m^*-j}) \\ = l_{m^*} + s\pi_{m^*} \sum_{j=1}^{U} j \left[\prod_{v=m^*}^{m^*+j-1} \gamma_v - \prod_{v=m^*-j}^{m^*-1} \gamma_v^{-1} \right] \end{array} , \tag{3.2}$$

Where $U = \min(m^* - 1, M - m^*)$. For any $j \leq U$, the terms in each pair may nearly cancel each other. Since the stationary distribution is peaked around target, the sum is dominated by the first $1 - 2$ summand pairs.[6] We now introduce Taylor expansions of F in the two levels adjacent to target:

$$F(l_{m^* \pm 1}) = F(l_{m^*}) \pm s f(l_{m^*}) + \frac{s^2}{2} f'(l_{m^*}) \pm \ldots , \tag{3.3}$$

[6] Clearly, if $U < 2$, i.e., if the target is too close to a boundary, there is substantial bias away from the boundary. Here this case will be neglected.

where f is the threshold pdf. For SU&D, the first summand pair in (3.2) then becomes

$$\begin{aligned} \gamma_{m^*} - \gamma_{m^*-1}^{-1} &= \tfrac{1}{2}\left[\tfrac{1}{F(l_{m^*+1})} - \tfrac{1}{1-F(l_{m^*-1})}\right] \\ &\cong -2f'(l_{m^*})s^2 \end{aligned} \quad , \tag{3.4}$$

a bias that is second-order in s with a sign opposite that of $f'(F_{0.5}^{-1})$ (for unimodal or amodal threshold distributions, this means a bias in the direction of F's skew). It is quite straightforward to show that all summand pairs in (3.2) for SU&D give negative second-order terms when thus expanded. Therefore, SU&D's stationary mean is a biased estimator of the median for asymmetric distributions, but this bias can be kept very small under reasonably fine spacing. The same analysis for BCD yields

$$\begin{aligned} \gamma_{m^*} - \gamma_{m^*-1}^{-1} &= \tfrac{\Gamma}{F(l_{m^*+1})} - \tfrac{1-\Gamma}{1-F(l_{m^*-1})} \\ &\cong \tfrac{2\Gamma-1}{\Gamma(1-\Gamma)} f(l_{m^*})s \end{aligned} \quad , \tag{3.5}$$

and for KR

$$\begin{aligned} \gamma_{m^*} - \gamma_{m^*-1}^{-1} &= p\left[\tfrac{1}{F(l_{m^*+1})} - \tfrac{1-\left(1-F(l_{m^*-1})\right)^k}{F(l_{m^*-1})\left(1-F(l_{m^*-1})\right)^k}\right] \\ &\cong 2\tfrac{(k+1)p-1}{p(1-p)} f(l_{m^*})s \end{aligned} \quad , \tag{3.6}$$

where p was used a a shorthand for the target (2.9). In both cases there is a downward first-order bias. For any $k > 1$ the BCD bias term (all other things being equal) is larger, by a factor of about 5/3 (the exact ratio as a function of k is found by substituting (2.9) for p, Γ). This explains the numerical observations of Gezmu (1996), who noted that KR is 'better centered' on target than same-target BCD; *cf.* also Fig. 1. GU&D$_{(0,1,k)}$ 'inherits' the properties of SU&D, meaning that its bias is only second-order and therefore usually smaller than that of the other two methods. For symmetric or upward-skewed threshold distributions (a realistic assumption when dealing with positive thresholds),

GU&D$_{(0,1,k)}$'s bias is positive - again, in line with numerical observations such as Fig. 1. Numerical results indicate that for all three methods, stationary biases are quite moderate except when using very coarse spacing.

4 Convergence Rates

4.1 Overview

For U&D methods, the term 'convergence rate' can be interpreted in a number of ways. In view of theoretical results about the stationary mode, one may be interested in the convergence of the probability that the empirical mode is the stationary mode, or more generally in the convergence of empirical frequencies to their stationary values. On the other hand, in view of the prevalence of averaging-based estimators, there is practical interest in the convergence of the empirical mean to the stationary mean. Fortunately, the three methods discussed here generate simple Markov chains with finite state spaces, and therefore all convergence rate comparisons should yield equivalent results. Here we focus on the empirical mean, which is the simplest, most intuitive, and carries a direct practical significance for most applications.

As Gezmu and Flournoy (2006) note, for finite-state Markov chains the exact treatment distribution after i trials can be calculated, given knowledge of initial conditions and of P, the design's transition probability matrix (tpm):

$$\rho^{(i)T} = \rho^{(1)T} P^{i-1}, \tag{4.1}$$

where $\rho^{(1)}$ is an initial probability vector over $\{l_m\}$ and P^{i-1} is P raised to the $i-1$-th power. As $i \longrightarrow \infty$, $\rho^{(i)} \longrightarrow \pi$, regardless of $\rho^{(1)}$. Convergence rates can be estimated using the tpm. The tpm's of the three Markovian methods discussed here are stochastic and irreducible. Hence, the real parts of their

eigenvalues are bounded between -1 and 1, and are usually labeled in decreasing order: $1 = \lambda_0 \geq Re(\lambda_1) \geq \ldots \geq Re(\lambda_{M-1}) \geq -1$. If the tpm is also reversible, that is $\pi_i P_{ij} = \pi_j P_{ji}\ \forall i, j$, then the eigenvalues are all real. In that case, it can be shown that the total variation distance from π converges as (Diaconis and Stroock, 1991)

$$\|P^N \rho^{(1)} - \pi\|_{var}^2 \propto \left\{ \sum_m \left| \rho_m^{(N+1)} - \pi_m \right| \right\}^2 \leq c\lambda_{\#}^{2N}, \tag{4.2}$$

where c is a constant and $\lambda_{\#} \equiv max\,(\lambda_1, \mid \lambda_{M-1} \mid)$. Thus, the rate of convergence is governed by the second-largest eigenvalue. For finite-state Markov chains, it is straightforward to show that the empirical mean would converge at half the rate of the total variation distance.

Unfortunately, comparing the 3 methods cannot be trivially reduced to an eigenvalue problem: KR tpm's (listing all internal sub-states as separate states) are $Mk \times Mk$ irreversible matrices. Moreover, all sub-states of the same external KR level actually represent the same experimental treatment. Therefore, it may make sense to compare convergence by 'marginalizing' KR to a $k \times k$ reversible matrix using (2.6). This is analogous to assuming that in KR experiments internal-state balance is achieved much faster than between-level balance.

4.2 Numerical Study

In order to directly compare convergence rates, an 'all other things being equal' numerical time-progression was performed, as in (4.1).[7] A summary of results for convergence 'upwards' from l_1 and some targets appears in Table 1 ('down-

[7] Two sets of initial conditions were examined: beginning at l_1 (similar to typical Phase I conditions) or at l_M, over a variety of right-skewed, left-skewed and symmetric threshold distributions, 3 values of s and $k = 1, 2, 3$. For each design scenario, 1000 parallel sets of thresholds were generated, and ensemble grand means of ρ at each trial i were calculated. Convergence rates were estimated via an exponential fit on the difference between the ensemble grand mean at time i and the method's stationary mean, calculated exactly using the formulae for π. All simulations were performed in R (R Development Core Team, 2005).

ward' convergence was faster in most scenarios, but the overall pattern was not substantially different; complete results are available from the author). Shown are the number of trials needed for the ensemble mean of ρ to converge 99% of the way from l_1 to the stationary mean. These numbers can be seen as a practical cutoff, beyond which the treatment chain is quasistationary, or 'as good as' a sample from π (one reason for choosing 99% is because halving the numbers on Table 1 conveniently yields the number of trials needed to converge a more lenient 90% of the way). Some observations from Table 1:

1. SU&D converges faster than any non-median method examined. This is related to the accepted assumption that convergence slows down as the target gets farther from the median.

2. For a given target p, convergence rates are most strongly affected by a combination of distribution properties and spacing - specifically, by the difference in F values between adjacent levels around target. For example, the 'shallowest' conditions (exponential thresholds and $M = 10$) cause the slowest convergence, and vice versa.

3. Among the three non-median methods, KR is solidly ahead, while BCD is slowest. Based on these and other numerical runs not shown here, one could expect BCD to take around $20 - 40\%$ more trials than same-target KR to reach the same degree of convergence. The performance gap increases with k.

4. The number of 'trials to 99% convergence' can be quite substantial when compared with the design limitations of applications such as Phase I clinical trials. Moreover, the Markovian method identical to '3+3's initial stage ($\text{GU\&D}_{(0,2,3)}$) converges very slowly - even slower than BCD. It takes $\text{GU\&D}_{(0,2,3)}$ $50 - 90\%$ more trials than KR ($k = 2$) to achieve the

Table 1: Comparative summary of convergence calculations for several design scenarios and targets. Shown are the number of trials needed to achieve 99% convergence of the ensemble grand mean of $\rho^{(N)}$ (the treatment probability distribution at time N) to the stationary mean, beginning with the entire probability mass on l_1 at trial 1. Numbers were rounded up to the nearest integer, except for GU&D, for which they were rounded up to the nearest multiple of k. The adjectives 'tight' and 'disperse' for the logistic and log-normal scenarios indicate smaller or larger scale parameters, respectively.

Distribution	Weibull (Shape Parameter)				Logistic		Log-Normal	
	1 (Exp.)	1.5	3.5	5	'Tight'	'Disperse'	'Tight'	'Disperse'
$M = 5$, Upward	**'Trials to 99% Stationarity'**							
SU&D	12	9	6	10	7	9	6	9
BCD, $p = 0.293$	19	17	15	17	14	17	13	16
KR, $k = 2$	**17**	**14**	**10**	**10**	**10**	**14**	**8**	**13**
GU&D $(0, 1, 2)$	20	16	12	**10**	12	16	12	16
BCD, $p = 0.347$	**18**	**15**	**12**	**12**	**12**	**15**	**10**	**14**
GU&D $(0, 2, 3)$	36	27	15	15	15	24	15	24
$M = 10$, Upward	**'Trials to 99% Stationarity'**							
SU&D	34	26	14	12	13	22	14	26
BCD, $p = 0.293$	44	38	30	31	28	39	23	33
KR, $k = 2$	**39**	**32**	**21**	**21**	**20**	**32**	**17**	**27**
GU&D $(0, 1, 2)$	46	36	24	22	22	36	20	32
BCD, $p = 0.347$	**44**	**35**	**24**	**23**	**22**	**34**	**19**	**31**
GU&D $(0, 2, 3)$	81	57	33	30	33	54	27	51

same degree of convergence, even though the former's target is closer to the median.

Table 1's convergence rate estimates are in rough agreement with eigenvalue predictions for BCD and GU&D (data not shown). KR rates show an agreement with the 'marginalized', reversible tpm eigenvalues, indicating that indeed internal-state balance happens fast for this method. The practical meaning of the numerical study's results is discussed below.

5 Discussion and Conclusions

5.1 Method Properties and Comparison

This paper provides proofs and other evidence corroborating recent numerical finds (Gezmu, 1996; Ivanova et al., 2003): BCD is inferior to the more commonly used but less studied KR in all practical aspects, namely bias of the stationary mean, concentration of treatments around target, and convergence rate. Additionally, KR is easier to administer (requiring no real-time randomization) and has better ethical properties for medical and toxicity applications, because at least k subjects are treated before each dose escalation. The only advantages of BCD compared with KR, are its ability to target any percentile (an advantage diminished by the lucky coincidence that $k = 2, 3$ provide targets in close proximity to $F^{-1}(0.3), F^{-1}(0.2)$, respectively), and its easier-to-analyze theoretical properties. But for the practicing researcher, theoretical elegance is no match to operating characteristics.

Between KR and GU&D$_{(0,1,k)}$, the latter has smaller stationary bias but the former converges more quickly. Whenever overall experiment duration is a more urgent constraint than the number of trials, or in applications where cohorts are typically used, GU&D$_{(0,1,k)}$ would be the best choice of the three. The U&D

method most closely resembling the '3+3' Phase I method, GU&D$_{(0,2,3)}$, was found numerically to converge very slowly. In view of the numbers on Table 1 it seems that under Phase I sample-size limitations the '3+3' method would quite often fail to converge. it is likely that KR with $k = 2$ or GU&D$_{(0,1,2)}$, coupled with adequate stopping rules and simple, optimized estimation, would provide a much better alternative for the task of finding the 30th percentile under these constraints.

5.2 Implications for Design

It is tempting to interpret Table 1 optimistically. Halving the numbers yields an estimate of 'trials to 90% convergence', which researchers may contend is good enough. Then, for KR or GU&D$_{(0,1,k)}$, one may assume quasi-stationarity within $5-20$ trials (depending upon F and design spacing). This may seem satisfactory, unless sample size is severely constrained - as is, unfortunately, quite often the case for Phase I experiments. Barring a lucky choice of starting point, the above-quoted number of trials should be perceived as a direct deduction from the effective sample size. We should also keep in mind that estimation uses either $\hat{F}$ (the binomial estimates of F at design points), or some average of treatments - which are, of course, a strongly dependent sample. Moreover, even under stationarity U&D treatments still spread over several levels around target. To sum it up, if one plans to use < 30 trials for a complete non-median-target U&D experiment, success is far from guaranteed. In a similar vein, the tendency to perform SU&D with $N < 10$ - existing since the 1950's and still prevalent in many fields - is statistically unfounded; the convergence rates in Table 1 indicate at least $15 - 20$ trials for reasonable median estimation. More generally, I suggest viewing U&D treatment chains as an unknown mix of nonstationary and quasistationary phases, and plan the design and estimation accordingly.

As shown on Section 4, very coarse U&D designs mitigate the convergence problem. However, as the preceding sections show, this comes at a price:

- A basic advantage of U&D - a steep stationary peak around target - is not realized if the design is too coarse. Moreover, unless there are at least $2-3$ levels between the target and either boundary, average-based estimators will be strongly biased away from the boundary.

- In cases when *'yes'* means an adverse response (e.g. toxicity), a coarse design puts considerably more subjects at risk of treatments far above target, even under stationarity (*cf.* Fig. 1).

- Neglecting boundary effects, the stationary mean's bias is roughly proportional to s or to s^2. This means that as s increases, the method's effective target (represented by the stationary mean) migrates away from $F^{-1}(p)$ and average-based estimators' performance deteriorates.

This convergence-precision tradeoff was noticed as early as Wetherill (1963); however, many researchers still follow Dixon and Mood (1948)'s original rules of thumb ($s \approx 2\sigma/3$ to $3\sigma/2$, under a normality assumption), which lead to effectively only $4-6$ levels. In order to utilize U&D's stationary advantages while keeping a reasonable convergence rate, I suggest dividing the range expected to contain the middle ~ 90 percentiles into $8-10$ levels.

Two-stage U&D designs, composed of a fast-converging start-up stage followed by a stage with better stationary properties, are quite common (Wetherill, 1963; Storer, 1989; Garcìa-Perez, 1998; Storer, 2001; Potter, 2002). The problem with such designs is that the number of trials to transition point is random. For the most common choice of a transition event - the first change in response type - this number follows a generalized geometric distribution, having a rather large variance. Additionally, some two-stage designs use an initial SU&D stage

followed by a stage targeting a lower percentile (Storer, 1989), meaning that the transition point is by definition off-target. In both cases, one may end up worse off with the start-up stage than without it, as Garcìa-Perez (1998) found in his numerical study.

5.3 Estimation

This article avoids discussing the final phase of U&D experiments - quantile estimation - and deals exclusively with convergence and stationary properties. Estimation has been studied by the author, and the current methods can be improved upon (see Oron (2005) for some preliminary results). However, in the rule-based U&D framework this can be considered a separate topic, because the U&D experiment may be followed by any estimation procedure (and indeed, a variety of estimation methods is encountered in practice). Since the U&D design having the best convergence and stationary properties (all other things being equal) would provide the most information about the target region, it should also yield the best estimation performance regardless of the chosen estimator.

Acknowledgments

This paper has evolved from a fall 2003 session of the University of Washington statistics and biostatistics departments' consulting service (a session initiated by anesthesiologist M. J. Souter), into Ph.D. dissertation work. Besides many members of the two departments, and other colleagues who have helped along the way, the author extends a special thanks to committee chair P. Hoff, to consulting director P. Sampson, and to N. Flournoy who provided crucial and timely comments.

References

Anderson, T., McCarthy, P., Tukey, J., 1946. staircase method of sensitivity testing. Naval Ordinanace Report 65-46, Statistical Research Group, Princeton University, Princeton, NJ, USA.

ASTM, 1991. Standard test method for estimating acute oral toxicity in rats. American Society for Testing and Materials, Philadelphia, USA, designation E 1163-90.

Bortot, P., Giovagnoli, A., 2005. Up-and-down experiments of first and second order. J. Statist. Plann. Inference 134 (1), 236–253.

Brownlee, K. A., Hodges, Jr., J. L., Rosenblatt, M., 1953. The up-and-down method with small samples. J. Amer. Statist. Assoc. 48, 262–277.

Camorcia, M., Capogna, G., Lyons, G., Columb, M., 2004. Epidural test dose with levobupivacaine and ropivacaine: determination of ED50 motor block after spinal administration. Brit. J. Anaesthes 92 (6), 850–853.

Capogna, G., Parpaglioni, R., Lyons, G., Columb, M., Celleno, D., 2001. Minimum analgesic dose of epidural sufentanil for first-stage labor analgesia. Anesthesiology 94, 740–744.

Chao, M. T., Fuh, C. D., 2001. Bootstrap methods for the up and down test on pyrotechnics sensitivity analysis. Statist. Sinica 11 (1), 1–21.

Chao, M. T., Fuh, C. D., 2003. Why wild values occur in pyrotechnic sensitivity analysis. Propell. Explos. Pyrotech. 28 (4), 216–218.

Derman, C., 1957. Non-parametric up-and-down experimentation. Ann. Math. Statist. 28, 795–798.

Diaconis, P., Stroock, D., 1991. Geometric bounds for eigenvalues of Markov chains. Ann. Appl. Probab. 1 (1), 36–61.

Dixon, W. J., 1965. The up-and-down method for small samples. J. Amer. Statist. Assoc. 60, 967–978.

Dixon, W. J., Mood, A., 1948. A method for obtaining and analyzing sensitivity data. J. Amer. Statist. Assoc. 43, 109–126.

Drover, D., Litalien, C., Wellis, V., Shafer, S., Hammer, G., 2004. Determination of the oharmacodynamic interaction of propofol and remifentanil during esophagogastroduodenoscopy in children. Anesthesiology 100, 1382–1386.

Durham, S. D., Flournoy, N., 1994. Random walks for quantile estimation. In: Statistical decision theory and related topics, V (West Lafayette, IN, 1992). Springer, New York, pp. 467–476.

Durham, S. D., Flournoy, N., 1995. Up-and-down designs. I. Stationary treatment distributions. In: Adaptive designs (South Hadley, MA, 1992). Vol. 25 of IMS Lecture Notes Monogr. Ser. Inst. Math. Statist., Hayward, CA, pp. 139–157.

Durham, S. D., Flournoy, N., Montazer-Haghighi, A. A., 1995. Up-and-down designs. II. Exact treatment moments. In: Adaptive designs (South Hadley, MA, 1992). Vol. 25 of IMS Lecture Notes Monogr. Ser. Inst. Math. Statist., Hayward, CA, pp. 158–178.

Durham, S. D., Flournoy, N., Rosenberger, W., 1997. A random walk rule for phase I clinical trials. Biometrics 53 (2), 745–760.

Garcìa-Perez, M., 1998. Forced-choice staircases with fixed step sizes: asymptotic and small-sample properties. Vision Res. 38, 1861–1881.

Gezmu, M., 1996. The geometric up-and-down design for allocating dosage levels. Ph.D. thesis, American University, Washington, DC, USA.

Gezmu, M., Flournoy, N., 2006. Group up-and-down designs for dose-finding. J. Statist. Plann. Inference 136 (6), 1749–1764.

Giovagnoli, A., Pintacuda, N., 1998. Properties of frequency distributions induced by general "up-and-down" methods for estimating quantiles. J. Statist. Plann. Inference 74 (1), 51–63.

Goodman, S., Zahurak, M., Piantadosi, S., 1995. Some practical improvements in the continual reassessment method for phase I studies. Stat. Med. 14, 1149–1161.

Hughes, B. D., 1995. Random Walks and Random Environments. Vol. 1. Oxford Science Publications. The Clarendon Press Oxford University Press, New York, random walks.

Ivanova, A., Flournoy, N., 2006. Up-and-down designs in toxicity studies. In: Statistical Methods for Dose-finding Experiments. Statistics in Practice. John Wiley & Sons, Chichester, pp. 115–130, ed: S. Chevret.

Ivanova, A., Haghighi, A., Mohanty, S., Durham, S., 2003. Improved up-and-down designs for phase I trials. Stat. Med. 22, 69–82.

Komori, Y., Hirose, H., 2000. An easy parameter estimation by the EM algorithm in the new up-and-down method. IEEE Trans. Dielect. Elect. Insul. 7 (6).

Lagoda, T., Sonsino, C., 2004. Comparison of different methods for presenting variable amplitude loading fatigue results. Mat. Wissen. Werkstoff 35 (1), 13–20.

Lichtman, A., 1998. The up-and-down method substantially reduces the number of animals required to determine antinociceptive ED50 values. J. Pharma. and Toxic. Meth. 40 (2), 81–85.

Marvit, P., Florentine, M., Buus, S., 2003. A comparison of psychophysical procedures for level-discrimination thresholds. J. Acoust. Soc. Am. 113 (6), 3348–3361.

NIEHS, 2001. The Revised Up-and-Down Procedure: A Test Method for Determining the Acute Oral Toxicity of Chemicals. National Institute of Environmental Health Sciences, National Institute of Health, Washington D.C., USA, NIH publication No. 02-4501.

OECD, 1998. The Revised Up-and-Down Procedure: A Test Method for Determining the Acute Oral Toxicity of Chemicals. Organisation for Economic Co-operation and Development, Paris, France.

O'Quigley, J., Pepe, M., Fisher, L., 1990. Continual reassessment method: a practical design for phase 1 clinical trials in cancer. Biometrics 46 (1), 33–48.

Oron, A., 2004. Letter to M.J. Souter, MD. Consulting summary report, Statistical Consulting Unit, Department of Statistics, University of Washington, Seattle, WA, USA, available upon request from department.

Oron, A., 2005. The up-and-down experimental method: Stochastic properties and estimators. Dependent data preliminary exam report, Statistics Department, University of Washington, Seattle, WA, USA, revised version, March 2005; available from the author.

Potter, D. M., 2002. Adaptive dose finding for phase I clinical trials of drugs used for chemotherapy of cancer. Stat. Med. 21, 1805–1823.

R Development Core Team, 2005. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0.
URL `http://www.R-project.org`

Rosenberger, W., Grill, S., 1997. A sequential design for psychophysical experiments: An application to estimating timing of sensory events. Stat. Med. 16, 2245–2260.

Rosenberger, W., Haines, L., 2002. Competing designs for phase I clinical trials: a review. Stat. Med. 21, 2757–2770.

Scherrer, S., Wiskott, A., Coto-Hunziker, V., Belser, U., 2003. Monotonic flexure and fatigue strength of composites for provisional and definitive restoration. J. Prosth. Dentist. 89 (6).

Storer, B. E., 1989. Design and analysis of phase I clinical trials. Biometrics 45 (3), 925–937.

Storer, B. E., 2001. An evaluation of phase I clinical trial designs in the continuous dose-response setting. Stat. Med. 20, 2399–2408.

Stylianou, M., Flournoy, N., 2002. Dose finding using the biased coin up-and-down design and isotonic regression. Biometrics 58 (1), 171–177.

Stylianou, M., Proschan, M., Flournoy, N., 2003. Estimating the probability of toxicity at the target dose following an up-and-down design. Stat. Med. 22 (4), 535–543.

Sunderam, R., Patra, R., Julli, M., Warne, M., 2004. Use of the up-and-down acute toxicity test procedure to generate LC50 data for fish. Bull. Environ. Contam. Toxic. 72 (5).

Treutwein, B., 1995. Minireview: adaptive psychophysical procedures. Vision Res. 35, 2503–2522.

Tsutakawa, R., 1967. Random walk design in bio-assay. J. Amer. Statist. Assoc. 62, 842–856.

von Bekesy, G., 1947. A new audiometer. Acta Oto.Laryn. 35, 411–422.

Weiss, G. H., 1994. Aspects and applications of the random walk. Random Materials and Processes. North-Holland Publishing Co., Amsterdam.

Wetherill, G. B., 1963. Sequential estimation of quantal response curves. J. Royal. Stat. Soc. B 25, 1–48.

Wetherill, G. B., Chen, H., Vasudeva, R. B., 1966. Sequential estimation of quantal response curves: A new method of estimation. Biometrika 53, 439–454.

Wetherill, G. B., Levitt, H., 1966. Sequential estimation of on a psychometric function. Brit. J. Math. Stat. Psych. 18, 1–10.