Visual Psychophysics and Medical Imaging: Nonparametric Adaptive Method for Rapid Threshold Estimation in Sensitivity Experiments

VICTOR KLYMENKO, STEPHEN M. PIZER, AND R. E. JOHNSTON

Abstract—**Rapid advances in biomedical imaging, including new technologies in image acquisition, reconstruction and display, and new algorithms in digital image processing, have generated the need for valid, reliable, efficient, and standardized psychophysical methods of observer performance evaluation, particularly with respect to observer sensitivity to imaging parameters. In recent years there has been considerable activity in the development of new psychophysical techniques measuring observer sensitivity, each of them with relative advantages and disadvantages in terms of efficiency, reliability, underlying statistical assumptions, computational simplicity and domain of application [1]-[74]. In this paper we describe the *m*-AFC transformed up-down adaptive method for the rapid determination of visual thresholds [28], [32], [33], [49], [50], [55] in medical images. The method is very efficient in obtaining thresholds to medical imaging parameters; in addition it is free from criterion bias, an important concern in radiology, and it is free from parametric assumptions about the stimulus scale, which are often unknown due to the complexity of medical images. We report issues of experimental design that arise in the use of this method and note the psychological caveats which should be followed with human observers. We present two experiments, where we demonstrate the method's efficacy in determining thresholds and psychometric functions in medical images.**

I. Introduction

Advances in medical imaging such as new technologies in image acquisition, reconstruction, display, and new algorithms in digital image processing, have generated the need for observer performance evaluation of a large number of imaging parameters [75], [76]. Section II briefly overviews the perceptual issues relevant to medical imaging and reviews traditional methods of measuring visual sensitivity [15], [16], [18], [77], [78]. Section III describes an efficient adaptive method, the transformed up-down method [33], for use in the rapid determination of visual thresholds in medical images. Section IV reports human observer experiments demonstrating the method and Section V reports our conclusions on the efficacy of the method.

Our main purpose is to point out the relevance of a method of threshold estimation, originally imported from psycho-acoustics [33] and gaining wide currency in basic research in visual psychophysics [79]-[83], to applied questions in medical imaging. We briefly review current methodological thinking in psychophysics and argue for the wider use of the procedure known as the transformed up-down method to investigate issues in medical imaging. In addition, we call attention to frequently disregarded and potentially confounding psychological and methodological issues which need to be considered when obtaining empirical thresholds to medical imaging parameters. The method we describe is useful in measuring visual sensitivity, when in the context of an experiment the image parameter of interest can be modified on-line by a computer controlled display.

II. Psychometric Functions and Medical Imaging

The quality of the medical image a radiologist sees, is the final product of a complex chain of transformations, beginning with the physical properties of the object and ending with the perceptual response of the observer [84], [85], what we call the object to observer pipeline (OTOP). In medical image viewing situations, the observer may need to identify or classify a feature from among a set of potential features, such as the type of lesion present; discriminate a feature, such as the location of a stenosis; or detect a feature, such as the presence or absence of a nodule in a noisy image. The observer may need to make a determination as to diagnosis, such as that the patient has pneumonia; or treatment such as the location of a planned surgical path relative to a nerve or blood vessel. In each of these situations, the observer's ability to perceive the relevant information in the image will be based on a series of decisions made about parameters at different OTOP stages. For example, early stage OTOP parameter decisions might concern what amount of radioactive material needs to be used for the PET scan, or what are the parameter settings for an image reconstruction algorithm; an intermediate OTOP stage parameter decision might involve what degree of enhancement is best when applying an adjustable adaptive histogram equalization image process-

Manuscript received November 13, 1989; revised May 2, 1990. This work supported in part by NIH PO1-CA-47982.

V. Klymenko is with the Department Radiology, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599.

S. M. Pizer and R. E. Johnston are with the Departments of Computer Science and Radiology, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599.

IEEE Log Number 9037580.

ing algorithm [86], [87]; a later stage decision might concern the dynamic luminance range to which a display device should be set. Whatever the type of response required, OTOP parameter setting decisions will effect the observer's performance based on how the observer's sensitivity corresponds to changes in the OTOP parameter. The OTOP parameter need not necessarily be examined in terms of sensitivity to an intensive stimulus dimension such as luminance or contrast, but in terms of any perceivable dimension such as, for example, size, depth, or location.

An example of an OTOP parameter decision would be the degree of edge enhancement in an unsharp masking algorithm. This decision might effect, for example, the lesion contrast needed to detect a particular type of lesion in a medical image. The experimental conditions, the OTOP parameter, might be four settings of a single unsharp masking algorithm, or four different algorithms. Here we wish to determine what stimulus level, lesion contrast in this example, is needed to reach threshold for each of the four OTOP conditions. Above some stimulus level (high lesion contrast), where performance will be perfect, an experimental condition will be in the suprathreshold region; below some level (low lesion contrast) where performance will be at chance, it is in the subthreshold region. Between these levels (intermediate lesion contrasts), in the threshold region, performance will usually [32] vary monotonically from chance to perfect. These functions are variously known as sensitivity, psychophysical or psychometric functions. These psychometric functions are almost always monotonic, generally ogival, and often (particularly for intensive stimulus dimensions) Gaussian integral functions, or logistic or Weibull functions [21], [24], [38], [43], [65], [77], [88], [89]. Also the stimulus region they map onto, the threshold region, is of particular importance for a number of reasons. Because of both technological OTOP and human performance limitations, this is currently the maximum range of many medical images. The threshold region is also the only one of the three regions which allows us to determine quantitatively and objectively how perceptual performance per se varies in terms of an image parameter variation.

In addition, it is also the only region in which we can equalize the performance effects of different parameters, or scale the same parameter in psychophysically equal units, since points both in the perfect suprathreshold and chance subthreshold ranges are indeterminate. For example, consider the case where one wants to perceptually linearize the 256 driving levels of a display device [90]. That is one wants to set the physical luminances corresponding to each driving level such that a target in the image at say 5 driving levels above its background is equally detectable throughout the range of driving levels—for example, a target set at driving level 15 in a background set at driving level 10 would be as perceivable as a target set at 205 in a background set at 200. Here the procedure would be to define a point on the psychometric function, such as the 75% correct point as the threshold or just noticeable difference (JND) unit, and determine the luminance increment JND's for each driving level [84], [85], [90]. Deriving equal luminance increments by subjective scaling such as Steven's magnitude estimation procedure [78], would produce subjectively equal units of the stimulus scale, equal brightness units in this case, but would likely be inappropriate in terms of objective task-specific performance, such as detection or discrimination of a target. In the demonstration experiments, we will determine what are the mimimum increments in driving level units of a target above its noisy background which will achieve threshold of the target for a set of two dynamic luminance ranges of a CT image displayed on a video screen.

As the nature of thresholds and psychometric functions are well covered elsewhere [16], [18], [77], [78], we only briefly mention several pertinent notions. Historically the research conducted under the rubric of signal detection theory has shown that thresholds are not absolute and can be decomposed into perceptual sensitivity and cognitive decision components. The decision components can be controlled when using an alternative forced choice (AFC) procedure, and the more alternatives the more stable the resulting index of sensitivity [77]. If one wishes to compare the effect of different experimental conditions on threshold, one can simply record percent correct in an m-AFC procedure and compare the results, and given additional assumptions compare the results to other m-AFC experiments, where m is some number of alternatives [91]. These percent corrects, often converted to some common currency such as d', will give us different threshold points on the psychometric functions for the different experimental conditions.

If one wishes instead to equalize the different experimental conditions, that is to find the same threshold point, such as the 50% correct point, on the different psychometric functions, one would need to vary the stimulus magnitude for each experimental condition to reach these points. Traditionally this has been done by the method of adjustment, where the observer sets the threshold, and the method of limits, where the observer indicates his threshold by changing his response to ascending and descending stimulus magnitude sequences. Although efficient, both methods are confounded by subjective criterion bias. This bias has been avoided by using the less efficient method of constant stimuli [15], [18], [53]. For each experimental condition, this last method presents a large number of trials in random order at predetermined stimulus levels, records resulting percent corrects and interpolates the desired threshold point. The method of constant stimuli is inefficient in that many trials will be presented at some distance from the desired threshold points, and in order to interpolate, one needs to make parametric assumptions about the stimulus scale onto which the psychometric function is mapped.

In examining the perceptual effects of directly varying a parameter somewhere in the object to observation pipeline, in medical imaging contexts one often cannot make any assumptions regarding the underlying stimulus scale of the psychometric function [82], [92], [93]. Earlier, we gave the example of the lesion contrast needed for detection at each of four settings of an OTOP parameter, four degrees of edge enhancement in an unsharp masking algorithm. We might instead be interested in how perceptual performance maps directly onto varying the settings of an OTOP parameter. Consider, for example, the case where the experimenter is interested in investigating the effects of pixel histogram shape upon visual threshold for detecting the location of a tumor, for example, is the brain tumor in contact with the optic nerve? The OTOP parameter of interest here, the look-up-table (LUT) remapping function, will produce a hyperbolic histogram at one extreme, a flat histogram for an intermediate value where each pixel value in the range occurs an equal number of times in the image, and a U-shaped histogram at the other extreme parameter setting. We wish to derive a psychometric function for correct detection of location in terms of how observer sensitivity maps directly onto the variation in this OTOP parameter. In this case, rather than varying a perceptually well specified parameter such as contrast to derive contrast thresholds and psychometric functions for each OTOP parameter variation, the experimenter is directly varying the OTOP parameter to derive its psychometric function. In this example, as the shape of the pixel histogram changes, the signal-to-noise ratio (SNR), the conspicuity of the lesion, the average luminance increment of the lesion over its background, the global and local luminance and contrast in the image, etc., will change in complex and often unknown ways. Even if one knows the relevant stimulus scale, calculating the relevant changes induced by varying the histogram shape may be mathematically intractable. On the other hand, one may assume that, for example, the Gaussian integral [94] of the psychometric function varies against the histogram shape changing units which we have imposed, when in reality it varies against some more complex function [93] of the histogram shape changing units, in which case the interpolated threshold estimate will be based on the wrong parametric assumptions, the result being unpredictable in terms of bias and accuracy [82]. Many image manipulations defy simple parametrization of the relevant stimulus scale which may not even be known. Even in well researched simple stimulus domains, there are still debates concerning the correct units of the underlying stimulus scale [82].

In the transformed up-down method described in the next section, we do not need to make any parametric assumptions within the context of the experiment. However, if one decides such assumptions are safe and useful and wishes to use parametric data analysis techniques such as probit analysis [33], [95], one can still use the data collected by this method, with the advantage that the experiment will be more efficient than the method of constant stimuli in terms of the number of trials and the placement of stimuli [15], [32], [33], [82]. This method is one of a class of modern techniques known as adaptive procedures, which are those procedures where the stimulus level on each trial in an experiment is adaptively based on the observer's previous responses. The transformed up-down method, unlike many other procedures, requires no parametric assumptions and is relatively robust to time course effects, which are the drifts in threshold caused by various external factors. These external factors are an important aspect of the threshold estimation problem we discuss next.

Assume for the moment that we know the stimulus level of the 50% point on a psychometric function for an m-AFC paradigm and want to demonstrate this. Logically one might assume that we should present all the trials at the stimulus level corresponding to the 50% point. How many trials should we present at this stimulus level? The deviation of the empirical percent from the real percent correct will of course follow the sampling distribution based on the number of trials and the value of m. This is the foundation of the "wall of variability" beyond which we cannot go even in principle. Additional bricks to the wall will be unavoidably added by between-subject variability, and by external factors, which introduce additional variability into each individual observer's responses and corresponding threshold estimates. They may in fact change the threshold being estimated during the process of estimation, for example, by perceptual learning, habituation, and aftereffects, even development of superstitious response behavior, and so on [74]. The purpose of good experimental design is, of course, to reduce the confounding influence of external factors which, because of the inherently stochastic nature of all threshold estimation procedures, cause unique problems. Although these are attenuated by the efficiency of adaptive methods, they are never absolutely eliminated because of the constraints of human experimentation. These constraints include the temporal effects of the psychological variables of fatigue, perceptual learning, motivation, light and dark adaptation and so on [74], as well as the inherent variability in the psychometric function itself. Each additional observation taken to estimate a threshold will, in a fashion paralleling the Heisenberg uncertainty principle in physics, affect the threshold. The improvement in threshold estimation of running a thousand rather than a hundred trials, thus, may not be as straightforward as the mathematics would imply due to the increasingly uncontrolled influence of psychological variables such as, for instance, fatigue. At some indeterminant point in time uncontrolled attentional lapses due to fatigue and boredom will begin irreparably confounding the sensitivity data [20]-[23]. Thus chipping away at one brick in the wall of variability will cause another to fall into its place. Long runs of near threshold stimuli are especially disconcerting, because these stimuli by their nature impose the maximum perceptual and attentional performance requirements on observers [96].

While theoretically, due to the multiple trials needed, no technique can completely tease apart a threshold estimate from the various confounding time course factors (which may have opposing effects such as perceptual learning and stress), the transformed up-down method allows one to track changes in the overall influence of these external factors by obtaining ongoing threshold estimates. The main advantage of adaptive techniques in general is their efficiency in terms of the number of trials. The main advantages of the adaptive method we will describe is its computational simplicity, ability to track drifting thresholds, and its freedom from restrictive parametric assumptions about the underlying stimulus scale onto which the psychometric function is mapped.

III. Nonparametric Adaptive Threshold Estimation

The transformed up-down method has been described as a modified method of limits [33], in that the threshold estimate is based on mini-ascending and -descending sequences of stimulus level, and alternatively as a modified method of constant stimuli [15], in that the ongoing threshold estimate is adaptively bracketed.

A. Transformed Up-Down Method of Wetherill and Levitt

Assume we have prior knowledge of the general threshold region. Once we are in this neighborhood, the simplest stimulus transition or stepping rule is to raise the stimulus level for the next trial following an incorrect response, or lower the stimulus level following a correct response. This transition rule (Rule A) leads to what is known as the simple up-down or staircase method. In the resulting up and down sequence of stimulus levels, changes in the direction of stimulus levels are called turning points; upper turning points for a change from an ascending to descending series of levels, and lower turning points for the reverse. The stimulus value midway between a turning point and the preceeding stimulus level constitutes a local estimate. To avoid directional bias, local estimates associated with an equal number of upper and lower turning points are averaged. This is usually more simply described as averaging the turning points [15]. The average stimulus value of the two turning points (an upper and neighboring lower turning point) constitutes a mid-run estimate, and the average of all the mid-run estimates is operationally defined in this method as the threshold estimate of the 50% point. The mid-run estimates, thus provide a rough moving window of the threshold estimate. Changing the transition rule so that the stimulus level is now only decreased after two consecutive correct responses, but still increased after each incorrect response (Rule B), will result in a higher threshold estimate, the 70.7% threshold point. The transformed up-down method, worked out by Wetherill and others [28], [29], [31], [67]–[71], refers to the set of potential transition rules defining different threshold points [32], [33], [35]. The logic behind the transition rules is as follows. Each Rule X has an associated percentage of correct responses, x, such that a stimulus at the $x\%$ correct level has an equal probability of ascending or descending. Whether the stimulus level rises or falls, it will subsequently have a greater probability of reversing direction. Thus the stimulus levels will tend to converge onto and oscillate about the $x\%$ correct point. For example, modifying the transition rules so that the stimulus level is now only decreased after three consecutive correct responses, but still increased after each incorrect response (Rule C), will result in a higher threshold estimate, the 79.4% threshold point.

Because of the differing number of trials specifying different transition rules, only a few of the potential quantal threshold points are practical within the context of observer sensitivity experiments [32], [33]. However two points, such as those specified by Rules A and B, are sufficient to derive a practical estimate of two threshold points and the spread, or steepness, of a psychometric function. The spread here is operationally defined as the difference between threshold estimates obtained by Rules A and B. This difference will allow us to determine whether the psychometric function has a small spread, is steep, or a large spread, is shallow, that is, how sensitive it is to changes in the stimulus parameter. The steepness of the psychometric function is sometimes more conveniently specified in terms of slope which is simply the inverse of the spread.

In estimating thresholds from the data, methods other than the average mid-run estimate [3], [10], such as the median of the stimulus levels visited or maximum likelihood estimation should not change the estimates in terms of the experimental results [46], [55]. If the differences between experimental conditions are so small as to be affected by the method of estimation, the results should be interpreted with extreme caution. One may always theoretically refine one's estimate based on the parametric and other assumptions one is willing to make; however, this will generally be of little practical consequence in terms of the empirical conclusions one is likely to derive [39], [51]–[53]. In addition to the method of estimation, the operational definition of a threshold is based on a number of procedural factors including the value of m in m-AFC paradigms, the step size, the trial placement rules, and the percentage point being estimated. The legitimacy of comparing results across experimental paradigms by converting to a common currency, such as d' in signal detection theory, has been debated [7], [31], [55], [72], [77]; it will depend on a number of factors including the type of stimulus dimension under investigation [7], [26]. In the context of m-AFC experiments, m may refer to the number of alternative stimuli presented on each trial, only one of which contains the target to be detected; or m may refer to the number of target choices to be discriminated, only one of which is present on each trial. The value of m should be as large as is practical (preferably larger than 2 [28], [29], [46], [50]), and the threshold point being estimated must be larger than $100/m$ percent correct, which is chance guessing. To reduce variability, avoid estimating threshold points near the extremes of the psychometric function [38], [39].

In order to increase experimental efficiency, the threshold-estimation stage is usually preceeded by a range-location stage, described next.

B. Hybrid Adaptive Method

The purpose of the range-location stage is to first find the general region in which the desired threshold point lies. Thus, a large change in stimulus magnitude, or step size, is used, and this stage is terminated once the region is located. This two stage or hybrid approach has one main advantage. The number of trials is further reduced, thus increasing the efficiency, without introducing experimenter bias into the design [5]. The threshold estimation stage then uses a smaller step size based on the experimenter's desired threshold accuracy and other factors such as the equipment's resolution limitations. If only the second stage were employed, the experimenter might introduce bias into the experiment by estimating each threshold region himself and starting each experimental condition in each of their estimated regions, or having no experimenter bias by starting each condition at the same level, thus causing a loss of efficiency, particularly if a large number of steps are required to reach some of the threshold regions. Many algorithms have been developed [9], [15], [16], [23], [32], incorporating this two stage or hybrid approach, see review in [74]. For the range-location stage, we suggest using a modified method of limits in which the experiment is started above threshold and terminated at the first incorrect response in the down sequence. This stimulus level then becomes the first trial in the threshold estimation stage. In the following experiments, we use two down sequences for the first stage and start the second stage at the average of the two termination levels. This has the advantage of attenuating the loss of efficiency due to lapses or numerous chance corrects in the beginning of the experiment, but the disadvantage of introducing more trials in this stage. The reason we do not use an ascending as well as a descending series in this stage, as is normally done in the method of limits, is that the final average value of the two series would be more variable, on average further from the threshold region, due to the asymmetry in the two series in *m*-AFC paradigms. In the descending series the observer can be incorrect above threshold when he should have been correct only due to a lapse, which should be minimal at the start of the experiment. However in an ascending series, the observer by chance could be correct far below the threshold region.

Once in the threshold-estimation stage, the choice of step size is determined by the importance the experimenter assigns to the accuracy versus efficiency tradeoff. In some experimental contexts, efficiency and experimenter bias issues may necessitate several reductions in step size in the range-location stage or in some procedures in the threshold-estimation stage [79]. The main caveat with reductions in step size is that it should not be so rapid in terms of the number of trials that if the stimulus level inadvertently ends up far from the threshold by a string of lucky guesses or lapses, the experiment cannot recover efficiently. In the threshold-estimation stage the data will usually indicate by rapidly increasing mid-run estimates that the stimulus level is in the subthreshold region, or by unusually long descending sequences that the stimulus level is in the suprathreshold region. For the final threshold estimate, the first one, two, or three mid-run estimates are usually discarded to guard against the range-location stage overshooting or undershooting the threshold region.

Other adaptive methods do not separate the data analysis for threshold estimation and the trial placement rules. They operate by placing the stimulus level on each trial at the cumulative threshold estimate of the pooled data from all preceeding trials, e.g. [63]–[65]. These techniques have relative disadvantages including: one, they are often based on parametric assumptions; two, they need to place many trials in the disconcerting subthreshold region [20]–[23], [43], where the observer is responding at chance levels for long sequences of trials; three, if the threshold drifts or there are lapses or a string of lucky guesses, the final threshold estimate necessarily incorporates the propagated errors; and four, they sometimes can not terminate in a reasonable amount of time [20], [23], [55]. There is currently considerable activity in the development of threshold estimation methodologies. A new intermediate adaptive method recently developed [52], which appears potentially useful, awaits empirical testing. Next we demonstrate the transformed up-down method.

IV. Experimental Demonstrations

The utility of the *m*-AFC transformed up-down method was examined in two experiments. Experiment 1 demonstrated its effectiveness in obtaining comparative threshold estimates and comparative spread estimates of psychometric functions for two display conditions. The data were separately analyzed in terms of two OTOP parameters in order to demonstrate the conclusions which can be derived about parameter variation in different stages of the object to observation pipeline. Experiment 2 further demonstrated the efficiency, consistency and reliability of the estimation procedure.

A. Experiment 1: Comparative Threshold and Spread Estimates of Two Psychometric Functions

The effect of two display mappings, Full Dynamic Range and Compressed Dynamic Range, on the Upper (70.7%) and Lower (50%) thresholds of a notched target in a CT scan was estimated using a 4-AFC transformed up-down procedure. The four experimental conditions, two thresholds for each of the two dynamic ranges, allowed us to compare the thresholds and the spread of the psychometric functions for the two display conditions.

1) Method: We describe some aspects of the experiments in more detail than is usual in order to explicate methodological considerations.

a) Observers: There were five observers, each with normal, or corrected to normal, vision. They were two of the authors (VK and REJ), two departmental secretaries (AC and TB), and a biomedical engineering student (DG).

b) Equipment and Stimuli: A Comtal image processing system hosted by a VAX 11/730 computer presented the stimuli on a gray-scale video monitor and recorded the responses, which were made via a numeric keypad. The spatial and luminance resolution of the monitor respectively was 104 pixels per degree of visual angle, and 256 driving levels. The stimulus, a simulated CT scan in image memory, contained pixel values ranging from 0 to 255. After passing through a look-up-table, these were mapped onto the 0–255 driving level range of the display monitor. The luminance scale, measured in the center of a uniformly driven full screen, was approximately logarithmic (driving level—luminance in cd/m^2: 0—.017, 128—.55, 255—19.87).

The CT scan was generated by a computer image processing procedure which simulated the output of a reconstruction algorithm. The three stages of creating this CT scan involved the following: one, defining several regions of uniform pixel value corresponding to organs; two, blurring the edges to simulate the partial volume effect; and three, simulating the effect of uncorrelated Poisson noise in the projection profiles previous to tomographic reconstruction; resulting in an image with realistically correlated noise. For each experimental trial, one of four C-shaped targets, superimposed by pixel value addition, was present in the CT image (see Figs. 1 and 2).

The target was an 8 × 8 pixel square with a 4 × 4 pixel square section removed from the middle of either the upper, lower, right or left sides, forming a bi-symetrical C-shaped target with the open end facing up, down, right, or left. For each trial the target was presented in one of nine contigious pixel locations on the CT scan. These nine locations were defined with reference to a 3 by 3 array of adjacent pixel coordinates. The set of nine target locations, forming the possible backgrounds of the target, had been of uniform pixel value—135, seven and a half driving levels above the center of the range, —before the noise was added in the third stage of creating the CT scan. On each trial, one of 36 possible stimuli (4 targets × 9 locations) were randomly presented. Thus we used a single "frozen noise" image with positional variation of the target. The mean and standard deviation of the 100 pixel region which formed the possible backgrounds of the target was 134.88 and 21.23, respectively. The CT scan target combination presented to the observer was zoomed up by a 2 × 2 pixel replication; thus there were 52 enlarged pixels per degree of visual angle in the stimulus presented to the observer. Between trials a display showing the four target choices was presented for 5 s. The luminance, 4.5 cd/cm^2, of the intertrial display served to maintain a constant level of global light adaptation to avoid changing observer sensitivity during the course of the experiment. Observers were motivated to fixate the intertrial screen which contained the target choices and the feedback described in the next section.

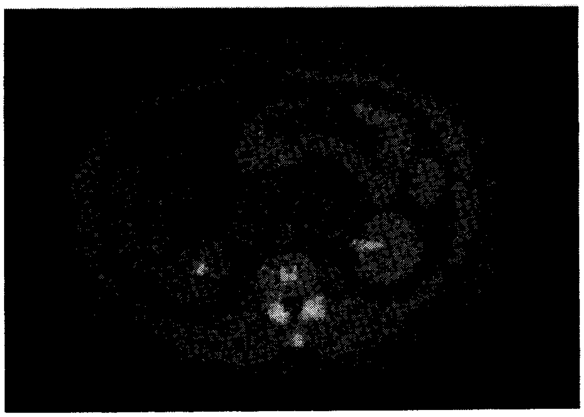

Fig. 1. Example of the stimuli presented in Experiments 1 and 2. CT scan displayed under Full Dynamic Range with a superimposed upward facing C-shaped target, one of the four alternative targets, see text.

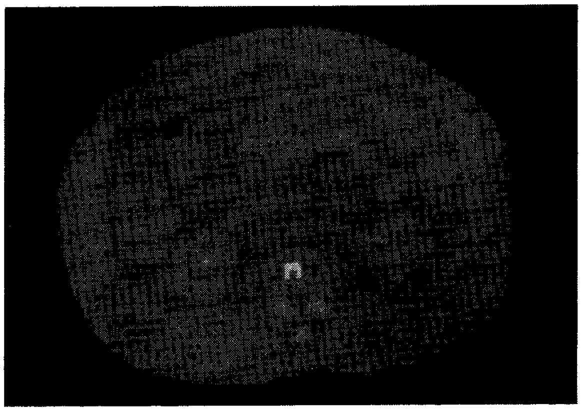

Fig. 2. Example of stimuli presented in Experiment 1. CT scan displayed under Compressed Dynamic Range with a superimposed downward facing C-shaped target, one of the four alternative targets, see text.

The DL value of the target in the CT scan was the number of driving level units the target pixels were raised above the noisy background. Each pixel in the target varied in absolute driving level, but all target pixels were a constant number of driving levels above the noisy background. The DL value of the target was manipulated adaptively as described in the next section.

In the Full Dynamic Range Displays, the 256 pixel value range of the CT scan in image memory was linearly mapped onto the full 256 driving level range of the display monitor by the identity LUT transform. In the Compressed Dynamic Range displays, the pixel value range was linearly mapped onto the central half of the driving level range. This produced a more faded image of lower contrast.

In describing the value of the target in the CT scan-target combination, we can refer to the physical DL units seen by the observer, or we can refer to the pixel values of the target in image memory before the dynamic luminance range compression, where we refer to the number of pixel values that the target is raised above the noisy background as "scaled CT" or SC units. Because of the LUT remapping function, in the Full Dynamic Range displays, DL units directly correspond to SC units, while in the Compressed Range displays, they are one half of the SC units. Because we have complete information, we will be able to determine which of these two OTOP parameters is more predictive of the threshold of the target.

c) Design: A version of the hybrid transformed up-down method described earlier was used to obtain the four threshold estimates—Upper and Lower Thresholds of the Full and Compressed Dynamic Ranges.

There were three stages presented in the following order in each experimental session: practice, range-location, and threshold-estimation. The type of feedback received was an observer's only indication of the stage. Some external factors such as fatigue and boredom can be partially ameliorated by motivating the observer with feedback.

In the practice stage the target value was always 60 DL units, far above threshold, and feedback was received after every trial in the form of the words "right" or "wrong" appearing after every trial on the intertrial display. There were 12 practice trials which consisted of the four experimental conditions being presented in random order within three blocks.

In the range-location stage, target values started at 60 DL for each of the four trial sequences representing the four experimental conditions. In each sequence every correct response decreased the target value by 6 DL units for the following trial in the sequence. The first incorrect response placed the target at 60 DL again, and the second incorrect response terminated the sequence. The four sequences were interleaved by presenting a trial from each sequence in random order for every block of four trials. As each sequence terminated, the remaining sequences were presented in random order in smaller blocks. Feedback was given after every trial in the form "correct" or "incorrect" appearing on the intertrial display.

The threshold-estimation stage began only after termination of all sequences in the range-location stage. In the threshold-estimation stage the starting value of the target of each sequence, representing each experimental condition, was the average value of the two incorrect responses recorded in the range-location stage. This value constitutes the initial estimate, near which the desired threshold is assumed to be located. The step size was one DL unit. The four sequences were interleaved in the same manner as in the range-location stage. In the threshold-estimation stage, feedback was given only after every four trials by presenting the number correct (0–4) on every fourth intertrial display. This decoupling of feedback from individual trials serves two purposes. One, along with interleaving, it prevents an observer from following and therefore being tempted to control the sequential dependencies in stimulus presentation [5], [32]–[34]. Two, it prevents "overlearning" unrelated to sensitivity. For example in our experiment, we used pseudorandom noise by having four target orientations in nine adjacent spatial locations creating a set of 36 target-noise combinations. If feedback was directly coupled to each trial, an observer might learn to attend to covarying features irrelevant to the task such as perceptually salient landmarks specific to certain target plus noise combinations. While it is unlikly that an observer would overlearn with 36 combinations in a short experiment, it is safer to decouple feedback from individual trials. In any case, as will be described below, one can check for time course effects such as overlearning.

The two Upper Threshold sequences followed Rule B defining the 70.7% point, and the two Lower Threshold sequences followed Rule A defining the 50% point. Each sequence was terminated after 14 turning points. For all sequences the turning point count started with the first upper turning point to avoid differential bias. If the first turning point was a lower turning point it was not counted. The average of each upper and following lower turning point constitutes a mid-run estimate. Thus each sequence produced 7 mid-run estimates. The first two mid-run estimates were discarded; thus for each observer, the final threshold estimate for an experimental condition was the average of the last five mid-run estimates.

d) Procedure: Each observer had experience with the experimental setup in similar experiments. Observation was binocular in a light-proof tunnel in a dark room. A chin rest positioned the observer 1.25 m from the display screen. Each observer dark adapted for ten minutes during which time the instructions were given. Each observer responded with the right hand via a standard numeric keypad which had a bump on the center key. Each observer was told to hit the key above, below, to the right, or to the left of the center key, respectively, depending upon whether the notch in the target faced up, down, to the right, or to the left, thus optimizing stimulus response compatability. Poor stimulus response compatibility, in requiring more attentional effort, will induce fatigue earlier and affect performance [96], [97]. The stimulus display remained on until the enter key was pressed at which time the response was registered. Responses could be changed before hitting the enter key as only the final response was registered. A short warning beep sounded if the observer hit a nonresponse key and a long beep sounded at the end of the experiment.

2) Results: For each observer, each experimental session lasted under one hour, with the following number of trials in the threshold-estimation stage: Full Dynamic Range, Upper Threshold— 54.0 (SD = 11.0), and Lower Threshold—32.8 (SD = 0.8); Compressed Dynamic Range, Upper Threshold—48.2 (SD = 4.3), and Lower Threshold, 30.2 (SD = 6.3). We were able to efficiently, without parametric assumptions, arrive at reliable estimates of the relative thresholds and spreads of the psychometric functions in two medical imaging display conditions in terms of two OTOP parameters. In addition we were able to track the time course of observer sensitivity in order to check for large confounding deviations from the psychometric stationarity assumption.

a) Threshold Estimates: Threshold-estimates, based on the average of the last five mid-run estimates, were obtained for each observer for each experimental condition. The data were separately examined in terms of two OTOP parameters: first, DL or driving level units, which map onto the physical luminance values of the target seen by the observer; and second, SC units, which correspond to prelook-up-table pixel values of the target in computer memory.

The four experimental conditions were the Upper and Lower Thresholds of the Full and Compressed Dynamic Ranges. In Table I the four means are given in DL units in the next to the last column, and in SC units in the last column.

Comparing the two threshold estimates within each dynamic range condition, we found a significant difference between the Upper and Lower Threshold of the Full Dynamic Range ($t(4) = 2.53, p < 0.05$), and between the Upper and Lower Threshold of the Compressed Dynamic Range ($t(4) = 2.87, p < 0.05$), indicating that, as expected, in both dynamic range conditions a greater increment in luminance above the background is needed to reach a higher point on the psychometric function. (The above t-tests are the same for DL and SC units.) No difference between Upper and Lower Threshold would have indicated either a very steep function, not resolvable in terms of the step size used in the experiment, and/or low statistical power in the experiment, indicating the need for more observers or more sessions per observer.

Comparing threshold estimates across dynamic range conditions, we found that the Upper Threshold of the Full Dynamic Range condition is greater than the Upper Threshold of the Compressed Dynamic Range condition in terms of DL units ($t(4) = 5.87, p < 0.01$), but not significantly different in terms of SC units ($t(4) = 1.31$). Similarly the Lower Threshold of the Full Dynamic Range condition is greater than the Lower Threshold of the Compressed Dynamic Range condition in terms of DL units ($t(4) = 2.88, p < 0.05$), but not significantly different in terms of SC units ($t(4) = 0.48$). The DL results indicate that as the physical luminance of the noise in the background is compressed, the luminance increment needed to correctly identify the target is also compressed. The SC results indicate that in terms of pixel values in image memory, the thresholds are the same in both dynamic range conditions.

As might be expected from the preceeding tests, when using all the available data (four threshold estimates) to test for an overall difference in location of the two psychometric functions, we found the following. In DL units, the difference in the average of the two threshold points for the Full Dynamic Range, 16.3 (SD = 1.6), and the Compressed Dynamic Range, 9.0 (SD = 1.5), was highly significant ($t(4) = 15.3, p < .01$). In SC units, the difference in the average of the two threshold points for the Full Dynamic Range, 16.3 (SD = 1.6), and the Compressed Dynamic Range, 18.1 (SD = 2.9), indicated that the location of the two psychometric functions are not significantly different, ($t(4) = 1.97$). The SC results show that in terms of pixel values in computer memory, the psychometric functions are in the same location in both dynamic range conditions.

The DL results show that the psychometric function for the Compressed Dynamic Range is shifted to a lower stimulus range. Of interest is whether this shift is uniform; that is, are the functions parallel [98]? In other words, is there a difference in the spread (Upper minus Lower Threshold point) of the Full Dynamic Range, 8.0 (SD = 7.1), and the spread of the Compressed Dynamic Range, 4.5 (SD = 3.5)? We found no significant difference, $t(4) = 1.91$, and thus we may assume, until we have evidence to the contrary, that in DL units, the two functions are close to parallel, at least within the power of the experiment. We may also assume that the family of psychometric functions between the full and compressed ranges are intermediate in thresholds and spreads. However, we would be hesitant to make any additional assumptions far outside this range. In general finding differences in spread (or slope) will require more statistical power than finding threshold differences. In SC units, we also failed to find any difference between the spread of the Full Dynamic Range, 8.0 (SD = 7.1), and the Compressed Dynamic Range, 9.1 (SD = 7.1), ($t(4) = 0.60$), further supporting the contention of no difference between the psychometric functions of the Full and Compressed Dynamic Range in terms of SC units.

For these medical images, the thresholds appear to be the same in SC units regardless of the physical luminances—the DL units—the observer sees. If for these noisy images, the two functions are the same, then the SNR in terms of SC units appears to be the limiting factor, independent of the dynamic luminance range of the display. If SC-SNR is invariant with respect to dynamic range, then we can conclude that our compression of the dynamic range of the display will have no effect on diagnosis. For example, a radiologist viewing the Compressed Dynamic Range will require a certain DL level of the target; increasing the dynamic range will only increase the DL level needed to see the target by an equivalent amount; performance is the same in terms of diagnosis. However, enhancing the target in terms of SC units, by adaptive histogram equalization for example, might enhance performance [87]. We explore this issue in more detail elsewhere [83].

Comparing the results of the SC and DL analysis, we can see that there is one inconsistency which has not been resolved, which is the failure to find a difference in spread of the two psychometric functions in terms of either units. Examining the data in terms of a difference in slope (the inverse of spread), we as expected, also found no difference in terms of either DL units, $t(4) = 1.47$, or SC units, $t(4) = 1.00$. Since the DL units of the Compressed Dynamic Range condition are a multiplicative constant

TABLE I
MEAN THRESHOLD ESTIMATES FOR EXPERIMENT I

	Mid-Run Estimates (DL Units)							Final Threshold Estimates	
Threshold Point Estimated	D1	D2	1	2	3	4	5	DL Units	SC Units
Full Dynamic Range									
Upper (70.7)	16.9 (7.1)	19.1 (5.1)	20.2 (4.6)	19.8 (4.3)	20.8 (4.8)	20.7 (4.3)	19.8 (4.7)	20.3 (4.2)	20.3 (4.2)
Lower (50)	14.5 (3.2)	13.3 (5.3)	13.4 (4.6)	13.2 (3.9)	12.1 (3.5)	11.3 (3.2)	11.2 (3.3)	12.2 (3.5)	12.2 (3.5)
Compressed Dynamic Range									
Upper (70.7)	9.4 (3.7)	10.3 (1.9)	10.2 (1.6)	11.1 (2.6)	11.7 (2.6)	11.9 (2.7)	11.6 (2.7)	11.3 (2.3)	22.6 (4.5)
Lower (50)	6.4 (4.1)	6.1 (3.3)	6.5 (3.1)	6.4 (3.0)	7.1 (2.4)	6.8 (2.7)	7.0 (2.5)	6.8 (2.4)	13.5 (4.7)
Mean of Four Points			12.6	12.6	12.9	12.7	12.4	12.7	

Note: D1 and D2 are the two discarded mid-run estimates. Final threshold estimates are the average of the last five mid-run estimates. Standard deviations in parenthess below the means are the between-subject standard deviations.

(0.5) of the SC units, the spread (and slope) of the Compressed Dynamic Range function must be different from the spread (and slope) of the Full Dynamic Range function in terms of at least one of the units. Thus if we were interested in these small differences, we could increase the statistical power of the experiment until one or the other (or possibly both) differences reached significance. In this experiment the LUT remapping function was very simple and we could examine the data in terms of both SC and DL units. However, there will be many cases in medical imaging in which an early OTOP parameter may not be so easily characterized in terms of the physical stimulus values presented to the observer.

b) Time Course of Threshold Estimates: Most techniques of threshold estimation implicitly or explicitly [65] assume stationarity of the psychometric function [20]; that is, there is no elevation or depression of a threshold over time due to cognitive factors such as overlearning, motivational factors, such as boredom, or perceptual-sensory factors, such as perceptual learning or light or dark adaptation. Table I shows the mean mid-run estimates; there are no general temporal trends. The bottom row of Table I shows the mean mid-run estimates averaged over the four conditions. We found no effect on thresholds of the ordinal position of these five (nondiscarded) mean mid-run estimates, $F(4, 16) = 0.21$. The results from the initial estimates are statistically indistinguishable from the final estimates. One still should obtain a number of mid-run estimates to enhance accuracy and avoid the undue influence of random statistical flucuations and lapses. (Lapses are a concern particularly if the stimulus duration is limited, where an observer needs to maintain a high degree of alertness.) In our experiment, we display the bright screen between trials to avoid contamination by dark adaptation. One might have performed a similar experiment without the bright intertrial screen if one were interested in studying the effects, if any, of the time course of dark adaptation on these thresholds. In such a case we would need to record the absolute times of mid-run estimates rather than merely the ordinal times. This adaptive method has been used to investigate the influence of various factors on the time course of sensory thresholds in tracking paradigms [73]. In the next experiment we examine the stationarity/time course issue in more detail.

B. Experiment 2: Reliability and Consistency of Threshold-Estimate

One of the experimental conditions tested in Experiment 1, the Upper Threshold of the Full Dynamic Range, was retested in Experiment 2, except this time five threshold estimates were obtained from each observer, allowing us to assess the time course and the internal consistency of the estimate as well as its reliability across experiments.

1) Method: There were four observers (VK, AC, TB, and DG). For each observer, there were five separate sequences following Rule B, for estimating the Upper Threshold point (70.7%) of the Full Dynamic Range condition. The sequences were presented in the same interleaved 4-AFC hybrid transformed up-down fashion described in Experiment 1, with four minor differences. One, there were ten rather than twelve practice trials in the practice stage. Two, for purposes of trial sequence randomization, a block of trials initially consisted of five rather than four trials; as before block size decreased as sequences terminated. Three, the step size was 10 DL units rather than 6 in the range-location stage. Four, feedback (0–5) was given after every five trials rather than every four trials in the threshold-estimation stage.

2) Results

a) Interexperiment Reliability: In Experiment 2, for each observer there were 25 mid-run estimates of the Upper Threshold point of the Full Dynamic Range, five from each sequence. The average of the 25, the mean Upper Threshold estimate was 20.6 (SD = 2.1). In Experiment 1, for the same four observers, where each observer's estimate was the average of five sequential mid-run estimates from a single sequence, the mean value was 19.7 (SD = 4.6). This difference, which is smaller than a step size, is not statistically significant $t(3) = 1.41$. Two conclusions can be derived from this. One, the method is reliable across experiments and minor design variations; and two, collecting approximately five times as much data did not significantly change the estimates or enhance the accuracy with respect to the conclusions of Experiment 1. Simultaneous interleaving of the different experimental conditions, as carried out in Experiment 1, is, however, still the best policy. It is doubtful that sophisticated parametric curve fitting or number crunching procedures would enhance the empirical data [25], [49], [50], particularly if one is trying to estimate a moving threshold. We examine this question next.

b) Time Course of Threshold Estimate: We analyzed the data in terms of its time course. In Experiment 1, each observer's threshold estimate for an experimental condition was the average of five temporally sequential mid-run estimates from one sequence. In Experiment 2, where five sequences tested the same condition, we examined the data in terms of the ordinal location of mid-run estimates. Each of the last five mid-run estimates was an experimental condition. In this case, an observer's threshold estimate was the average of five co-ordinal mid-run estimates from the five different sequences. For example, an observer's estimate of experimental condition 2 was the average of the 2nd mid-run estimate from each of the five independent sequences. Due to the nature of the design, coordinal mid-run estimates were not exactly in temporal phase. The means for Experiment 2 are shown in Table II. The means and between subject standard deviations for the same four observers from Experiment 1 are shown for comparison. Excluding the initial two discarded mid-run estimates, the means for mid-run estimates 1-5 were subjected to an analysis of variance test. The effect of experimental condition was not significant, $F(4, 12) = 1.57$, MSr = 1.36; however, the threshold estimate rises slightly, and we found that this slight linear trend reached significance, $F(1, 12) = 5.49, p < 0.05$. This small increase was probably due to occasional lapses from the increasingly unavoidable boredom of viewing the same stimulus display a large number of times. There are those who have argued that the last estimate, or even the stimulus levels of the last few trials, it the best estimate [3], [10]; however the same arguments concerning the undue influence of statistical fluctuation given earlier still apply.

c) Intraobserver Consistency: In testing aspects of medical images, we are generally not interested in individual differences where some observers may have generally higher thresholds due to opthalmological and other factors [19], but we are interested in how consistent the method is in measuring each individual observer's threshold. We examine this in terms of the variability of each individual observer's five threshold estimates as measured by the within-subject standard deviation of each observer's five mid-run estimates from independent sequences of trials. Row 5 of Table II lists the mean within-subject standard deviation for the five estimates for each of the seven ordinal positions in the stimulus sequences. The effect of the seven ordinal positions on the within-subject standard deviation of mid-run estimates was significant, $F(6, 18) = 5.02$, MS$r = 0.94$, $p < 0.01$. As discussed earlier we discard the first two mid-run estimates to reduce starting point bias in the final estimate. We found that within-subject standard deviation of the first two estimates was significantly larger than the last five estimates, $F(1, 18) = 27.82, p < 0.01$. As shown in Table II the within-subject standard deviations stabilize after the first two discarded mid-run estimates. It is unlikely that additional mid-run estimates would have any lower intraobserver variability. After the second discarded mid-run estimate, the independent mid-run estimates appear to run up against the wall of variability, akin to Planck's constant in physics, beyond which they cannot go due to various factors, most importantly the inherent variability of the psychometric function itself.

This last source of variability has been the subject of theoretical debates, where depending on the experimental conditions, it has been attributed to variability in the stimulus per se limiting ideal observers, or variability in the observer anywhere from the early visual neural system to late cognitive decision processes [78]. Our purpose is not to enter this debate, but merely to point out the utility of the m-AFC transformed up-down method as a practical and efficient procedure for reducing controllable external variance in estimating thresholds. Since we did not find any overlearning effect, it is possible that some of the within-subject variance was due to the stimulus variability generated by the nine positional perturbations of the target. If we were interested in separating the effect of stimulus variability from variability in the visual-brain system, we could have interleaved nine independent threshold estimation sequences, one for each position. Or if we were instead interested in generalizing to a larger set of stimuli, we could have instead randomly varied the noise and/or CT images instead of, or in additional to the positional perturbations. In this case, the resulting psychometric function would be a statistical composite of the individual psychometric functions. We did not find any overlearning in our experiments. However, the medical imaging investigator is cautioned that the need to check for this particular decremental time course effect is particularly pertinent when large numbers of trials are used with small stimulus sets. As shown in Table II the mean within-subject standard deviation (2.8) of the final threshold estimate is only slightly, but significantly, less than the average within-subject standard deviations of the five mid-run estimates (3.1); $t(3) = 5.20, p < 0.05$; thus indicating that, with this stimulus set we have clearly reached the point of diminshing returns in terms of the number of trials. Although there was slight incremental time course effect over the last five mid-run threshold estimates, it is interesting to note that it is unrelated to their consistency. In most medical imaging experiments where several conditions are tested, interleaving will generally eliminate the differential effect of external time course factors on threshold estimates from different experimental condi-

TABLE II
MEAN THRESHOLD ESTIMATES FOR EXPERIMENT 2

	Mid-Run Estimates							Final Threshold Estimate
	D1	D2	1	2	3	4	5	
Mean	17.6 (17.0)	19.3 (19.5)	19.9 (19.6)	20.2 (19.0)	20.4 (19.1)	20.7 (20.3)	21.8 (20.3)	20.6 (19.7)
Between Subject Standard Deviation	4.4 (8.2)	4.4 (5.8)	3.5 (5.1)	2.6 (4.6)	2.2 (3.4)	0.8 (4.8)	1.8 (5.3)	2.1 (4.6)
Mean Within Subject Standard Deviation	5.7	4.7	3.1	3.0	3.2	3.0	3.1	2.8

Note: Each mid-run and the final threshold estimate is the mean from five independent sequences of trials for four observers. D1 and D2 are the two discarded mid-run estimates. Final threshold estimate is the average of the last five mid-run estimates. Mean within-subject standard deviation is the standard deviation for each observer's five sequences averaged over the four observers. Shown for comparison in parentheses below the mean threshold estimates and below the between-subject standard deviations are the corresponding results from Experiment 1 for the same four observers for the same experimental condition.

tions. Where there are too many experimental conditions to administer in a single session, the researcher is cautioned to employ the appropriate counterbalancing [99] and replication of threshold estimates.

In order to more fully exploit the efficiency of adaptive procedures, a more realistic experiment than our demonstration experiments might have examined the effects of a larger number of experimental conditions (i.e., a large number of degrees of compression of dynamic ranges) on threshold estimates [83]. The results of such an experiment would be plotted as a threshold curve, one point representing a threshold estimate for each experimental condition. Corwin [100] recently has developed a highly efficient procedure for exploiting knowledge of previous threshold estimates to obtain new threshold estimates. In Experiment 1, if we had a third experimental condition, an Intermediate Dynamic Range between the Full and Compressed Dynamic Ranges, we might have initially searched for its threshold at an interpolated stimulus level between the previously obtained threshold estimates for the Full and Compressed Dynamic Ranges. This is most appropriate for small changes in the experimental variable as many threshold curves are nonmonotonic, the classic case being the spatial frequency contrast sensitivity curve [82]. Corwin's method [100] takes this into account by on-line varying not only the stimulus level, but also the degree of change from one experimental condition to the next based on the ongoing estimate of the local slope of the threshold curve.

V. CONCLUSION

Developements in threshold estimation continue unabated [100], [101], [102]. The numerous methods of threshold estimation in use and currently being developed and tested each have relative theoretical and empirical advantages and disadvantages. In terms of efficiency, reliability, and consistency, the transformed up-down method has been found to be equivalent to or compared favorably with other adaptive and nonadaptive methods in terms of computer simulations and empirical threshold estimates [31], [49], [50], [55], [72]. In addition to its efficiency, the transformed up-down method has qualitative advantages over other methods in that it can track drifting thresholds, and is free from parametric assumptions.

In Experiment 1, where two OTOP parameters were available for analysis, we were able to show that increasing the dynamic luminance range caused a corresponding increase in the thresholds in terms of DL units, but not in terms of SC units. We also showed that additional statistical power was needed in order to find a spread (or slope) difference, which logically must exist in terms of one of the two OTOP parameters. If only one OTOP parameter was available, one could not infer this, but would need to decide statistical power on the basis of how much of a difference is relevant to the medical imaging question under investigation. For example, the degree of precision required in calibrating an OTOP parameter might be decided on the basis of the steepness of the relevant psychometric function. The transformed up-down method was reliable across experiments and minor design variations. In Experiment 2, there was a slight time course effect. As we've shown, this method allows one a very straightforward manner of statistically checking for drifting thresholds. The interexperimental reliability results and the intraobserver consistency results both show that in estimating thresholds, the method rapidly converges down to the wall of variability inherent in the psychometric function. Increasing the amount of data collection (or number crunching sophistication) is always an option, but often likely to be an option of rapidly diminishing returns, due both to the negligible empirical differences in the obtained estimates and the increasing influence of external psychological variables.

We have found the hybrid m-AFC transformed up-down method to be a highly efficient, reliable, and practical procedure for the investigation of visual sensitivity issues in medical imaging.

ACKNOWLEDGMENT

We thank Dr. R. G. Swensson for a discerning critique of the manuscript.

REFERENCES

[1] I. Abramov, L. Hainline, J. Turkel, E. Lemerise, H. Smith, J. Gordon, and S. Petry, "Rocket-ship psychophysics: Assessing visual functioning in young children," *Investigative Ophthalmology & Visual Sci.*, vol. 25, pp. 1307-1315, 1984.

[2] D. Anbar, "The application of stochastic approximation methods to the bio-assay problem," *J. Statistical Planning Inference*, vol. 1, pp. 191-205, 1977.

[3] K. A. Brownlee, J. L. Hodges, and M. Rosenblatt, "The up-and-down method with small samples," *Amer. Statist. Assoc.*, vol. 48, pp. 262-277, 1953.

[4] R. A. Campbell and E. Z. Lasky, "Adaptive threshold procedures: BUDTIF," *J. Acoust. Society Amer.*, vol. 44, pp. 537-541, 1968.

[5] T. N. Cornsweet, "The staircase-method in psychophysics," *Amer. J. Psych.*, vol. 75, pp. 485-491, 1962.

[6] T. R. Corwin, R. T. Kintz, and W. J. Beaty, "Computer aided estimation of psychophysical thresholds by Wetherill tracking," *Behaviour Res. Methods Instrument.*, vol. 11, pp. 526-528, 1979.

[7] C. D. Creelman and N. A. Macmillan, "Auditory phase and frequency discrimination: A comparison of nine procedures," *J. Experimental Psych.: Human Percept. Perform.*, vol. 5, no. 1, pp. 146-156, 1979.

[8] C. D. Creelman and H. L. Kaplan, "Simultaneous independent threshold estimates: Multiple PEST," *Behavior Res. Methods & Instrument.*, vol. 5, pp. 89-92, 1973.

[9] C. Derman, "Non-parametric up-and-down experimentation," *Annals Mathematical Stat.*, vol. 28, pp. 795-798, 1957.

[10] W. J. Dixon and A. M. Mood, "A method for obtaining and analyzing sensitivity data," *J. Amer. Statist. Assoc.*, vol. 43, pp. 109-126, 1948.

[11] P. L. Emerson, "Observations on a maximum likelihood method of sequential threshold estimation and a simplified approximation," *Percept. Psychophys.*, vol. 36, pp. 199-203, 1984.

[12] —, "Observations on maximum-likelihood and Bayesian methods of forced-choice sequential threshold estimation," *Percept. Psychophys.*, vol. 39, pp. 151-153, 1986.

[13] —, "A quadrature method for Bayesian sequential threshold estimation," *Percept. Psychophys.*, vol. 39, pp. 381-383, 1986.

[14] P. L. Emerson and R. N. Sollod, "Are there any psychophysical applications of single-observation confidence intervals," *Percept. Psychophys.*, vol. 39, pp. 307-308, 1986.

[15] J. C. Falmagne, "Psychophysical measurement and theory," in *Handbook of Perception and Human Performance*, Vol. I, K. R. Boff, L. Kaufman, and J. P. Thomas, Eds. New York: Wiley, 1986, pp. 1-1-1-56.

[16] J. C. Falmagne, *Elements of Psychophysical Theory*. New York: Oxford, 1985.

[17] J. M. Findlay, "Estimates on probability functions: A more virulent PEST," *Percept. Psychophys*, vol. 23, pp. 181-185, 1978.

[18] G. A. Gescheider, *Psychophysics: Method and Theory*. Hillsdale, N.J.: Erlbaum, 1976.

[19] A. P. Ginsburg and M. W. Cannon, "Comparison of three methods for rapid determination of threshold contrast sensitivity," *Investigative Ophthamology Visual Sci.*, vol. 74, pp. 798-802, 1983.

[20] J. L. Hall, "A procedure for detecting variability of psychophysical thresholds," *J. Acoust. Soc. Am.*, vol. 73, no. 2, pp. 663-667, 1983.

[21] —, "Maximum-likelihood sequential procedure for estimation of psychometric functions," *J. Acoust. Soc. Am.*, vol. 44, pp. 370, 1968.

[22] —, "PEST: Note on the reduction of variance of threshold estimates," *J. Acoust. Soc. Am.*, vol. 55, pp. 1090-1091, 1974.

[23] —, "Hybrid adaptive procedure for estimation of psychometric functions," *J. Acoust. Soc. Am.*, vol. 69, pp. 1763-1769, 1981.

[24] L. O. Harvey, Jr., "Efficient estimation of sensory thresholds," *Behavior Res. Methods, Instruments, Comput.*, vol. 18, pp. 623-632, 1986.

[25] A. Hesse, "Comparison of several psychophysical procedures with respect to threshold estimates, reproducibility and efficency." *Acustica*, vol. 59, pp. 263-273, 1986.

[26] W. Jesteadt and R. C. Bilger, "Intensity and frequency discrimination in one-and two-interval paradigms," *J. Acoust. Soc. Amer.*, vol. 55, pp. 1266-1276, 1974.

[27] W. E. Kappauf, "Use of an on-line computer for psychophysical testing with the up-&-down method," *Amer. Psychol.*, vol. 24, pp. 207-211, 1969.

[28] C. D. Kershaw, "Statistical properties of staircase estimates from two interval forced choice experiments," *British J. Math. Stat. Psych.*, vol. 38, pp. 35-43, 1985.

[29] —, "Asymptotic properties of w, an estimator of the ED50 suggested for use in up-and-down experiments in bio-assay," *Ann. Stat.*, vol. 13, pp. 85-94, 1985.

[30] H. Kesten, "Accelerated stochastic approximation," *The Ann. Math. Stat.*, vol. 29, pp. 41-59, 1958.

[31] B. Kollmeier, R. H. Gilkey, and U. K. Sieben, "Adaptive staircase techniques in psychoacoustics: A comparison of human data and a mathematical model," *J. Acoust. Soc. Am.*, vol. 83, no. 5, pp. 1852-1861, 1988.

[32] H. Levitt, "Adaptive testing in audiology," in *Sensorineural Hearing Impairment and Hearing Aids, Scand. Audiol. Suppl.*, vol. 6, pp. 241-291, 1978.

[33] —, "Transformed up-down methods in psychoacoustics," *J. Acoust. Soc. Am.*, vol. 44, pp. 467-477, 1971.

[34] —, "Testing for sequential dependencies," *J. Acoust. Soc. Amer.*, vol. 43, pp. 65-69, 1968.

[35] H. Levitt and M. Treisman, "Control charts for sequential testing," *Psychometrika*, vol. 34, pp. 509-518, 1969.

[36] H. R. Lieberman and A. P. Pentland, "Microcomputer-based estimation of psychophysical thresholds: The best PEST," *Behavior Res. Methods Instrument.*, vol. 14, pp. 21-25, 1982.

[37] H. R. Lieberman, "Computation of psychophysical thresholds using the profit technique," *Behavior Res. Methods Instrument.*, vol. 15, pp. 446-448, 1983.

[38] S. P. McKee, S. A. Klein, and D. Y. Teller, "Statistical properties of forced choice psychometric functions: Implications of probit analysis," *Percept. Psychophys.*, vol. 37, pp. 286-298, 1985.

[39] R. Magidan and D. Willams, " Maximum-likelihood psychometric procedures in two-alternative forced-choice: Evaluation and recommendations," *Percept. psychophys.*, vol. 42, pp. 240-249, 1987.

[40] L. Marshall and W. Jesteadt, "Comparison of pure-tone audibility thresholds obtained with audiological and two-interval forced-choice procedures," *J. Speech Hear. Res.*, vol. 29, pp. 82-91, 1986.

[41] J. K. O'Regan and R. Humbert, "Estimating psychometric functions in forced-choice situations: Significant biases found in threshold and slope estimations when small samples are used," *Percept. Psychophys.*, vol. 46, pp. 434-442, 1989.

[42] M. Pavel, "A new adaptive method for forced-choice experiments," *J. Optical Soc. Amer.*, vol. 71, S48 (abstract), 1981.

[43] A. Pentland. "Maximum likelihood extimation: The best PEST," *Percept. Psychophys.*, vol. 28, pp. 377-379, 1980.

[44] I. Pollack, "Methodological determination of the PEST (Parameter Estimation by Sequential Testing)," *Percept. Psychophys.*, vol. 3, pp. 285-289, 1968.

[45] H. Robbins and S. Monro, "A stochastic approximation method," *Ann. Math. Statist.*, vol. 22, pp. 400-407, 1951.

[46] R. M. Rose, D. Y. Teller, and P. Rendleman, "Statistical properties of staircase estimates," *Percept. Psychophys.*, vol. 8., no. 4, pp. 199-204, 1970.

[47] J. Sacks, "Asymptotic distribution of stochastic approximation procedures, *Ann. Math. Stat.*, vol. 29, pp. 199-204, 1958.

[48] R. Sekuler and P. Tynan, "Rapid measurement of contrast-sensitivity functions," *Amer. J. Optometry Physiol. Optics*, vol. 54, pp. 573-575, 1977.

[49] B. R. Shelton, M. C. Picardi, and D. M. Green, "Comparison of three adaptive psychophysical procedures," *J. Acoust. Soc. Am.*, vol. 71, pp. 1527-1533, 1982.

[50] B. R. Shelton and I. Scarrow, "Two-alternative versus three-alternative procedures for threshold estimation," *Percept. Psychophys.*, vol. 35, no. 4, pp. 385-392, 1984.

[51] B. R. Shelton, "Rapid calculation procedures for the maximum-likelihood method of adaptive psychophysics," *Behavior Res. Methods Instrument.*, vol. 15, no. 1, pp. 87-88, 1983.

[52] W. A. Simpson, "The step method: A new adaptive psychophysical procedure," *Percept. Psychophys.*, vol. 45, no. 6, pp. 572-576, 1989.

[53] —, "The method of constant stimuli is efficient," *Percept. Psychophys.*, vol. 44, no. 5, pp. 433-436, 1988.

[54] C. W. Stucky, C. L. Hutton, and R. A. Campbell, "Decision rules in threshold determination," *J. Acoust. Soc. Amer.*, vol 40, pp. 1174-1179, 1966.

[55] J. A. Stillman, "A comparison of three adaptive psychophysical procedures using inexperienced listeners," *Percept. Psychophys.*, vol. 46, no. 4, pp. 345-350, 1989.

[56] M. M. Taylor and C. D. Creelman, "PEST: Efficient estimates on probability functions," *J. Acoust. Soc. Amer.*, vol. 41, pp. 782-788, 1967.

[57] M. M. Taylor, S. M. Forbes, and C. D. Creelman, "PEST reduces bias in forced choice psychophysics," *J. Acoust. Soc. Amer.*, vol. 74, pp. 1367-1374, 1983.

[58] M. M. Taylor, "On the efficiency of psychophysical measurement," *J. Acoust. Soc. Amer.*, vol. 49, pp. 505-508, 1971.

[59] R. K. Tsutakawa, "Random walk design in bio-assay," *J. Amer. Stat. Assoc.*, vol. 62, pp. 842-856, 1967.

[60] B. K. Tsutakawa, "Asymptotic properties of the block up-and-down method in bio-assay," *Ann. Statist.*, vol. 38, pp. 1822-1828, 1967.

[61] N. F. Viemeister, "Intensity discrimination: Performance in three paradigms," *Percept. Psychophys.*, vol. 8, pp. 417-419, 1970.

[62] W. S. Verplanck and J. W. Cotton, "The dependence of frequencies of seeing on procedural variables: 1. Direction and length of series of intensity-ordered stimuli," *J. Gen. Psych.*, vol. 53, pp. 37-47, 1955.

[63] A. Wald, *Sequential Analysis*. New York: Dover, 1947.

[64] A. B. Watson and D. G. Pelli, "The QUEST staircase procedure," *Applied Vision Association Newsletter*, vol. 14, pp. 6-7, 1979.

[65] A. B. Watson and D. G. Pelli, "QUEST: A Bayesian adaptive psychometric method," *Percept. Psychophys.*, vol. 33, pp. 113-120, 1983.

[66] R. J. Watt and D. P. Andrews, "A.P.E.: Adaptive probit estimation of psychometric functions," *Curr. Psych. Rev.*, vol. 1, pp. 205-214, 1981.

[67] G. B. Wetherill, "Sequential estimation of quantal response curves (with discussions)," *J. R. Statist. Soc. B*, vol. 25, pp. 1-48, 1963.

[68] G. B. Wetherill and H. Levitt, "Sequential estimation of points on a psychometric function," *Brit. J. of Math Stat. Psychol.*, vol. 18, pp. 1-10, 1965.

[69] G. B. Wetherill and H. Chen, "Sequential estimation of quantal response curves. II A new method of estimation," *Bell Telephone Laboratories Technical Memorandum*, 1964.

[70] G. B. Wetherill, *Sequential Methods in Statistics*. London, England: Chapman Press, 1971.

[71] G. B. Wetherill, H. Chen, and R. B. Vasudeva, "Sequential estimation of quantal response curves: A new method of estimation," *Biometrika*, vol. 53, pp. 439-454, 1966.

[72] C. C. Wier, W. Jesteadt, and D. M. Green, "A comparison of method of adjustment and forced choice procedures in frequency discrimination," *Percept. Psychophys.*, vol. 19, pp. 75-79, 1976.

[73] J. Zwislocki, F. Maire, A. S. Feldman, and H. Rubin, "On the effect of practice and motivation on the threshold of audibility," *J. Acoust. Soc. Amer.*, vol. 30, pp. 254-262, 1958.

[74] M. Zaus, "Hybrid adaptive methods," in *Progress in Mathematical Psychology I.*, -E. E. Roskam and R. Suck, Eds. Amsterdam, The Netherlands: North-Holland, 1987, pp. 351-378.

[75] H. K. Huang, *Elements of Digital Radiology*, Englewood Cliffs, NJ: Prentice Hall, 1987.

[76] R. A. Robb (Ed.) *Three-Dimensional Biomedical Imaging*, Vol. I and II. Boca Raton, Florida: CRC Press, 1985.

[77] D. M. Green and J. A. Swets, *Signal Detection Theory and Psychophysics*. New York: Wiley, 1966.

[78] J. C. Baird and E. Noma, *Fundamentals of Scaling and Psychophysics*. New York: Wiley, 1978.

[79] J. B. Calderone and M. K. Kaiser. "Visual acceleration detection: Effect of sign and motion orientation," *Percept. Psychophys.*, vol. 45, no. 4, pp. 391-394, 1989.

[80] T. Heckmann and E. M. Schor, "Panum's fusional area estimated with a criterion-free technique," *Percept. Psychophys.*, vol. 45, no. 4, pp. 297-306, 1989.

[81] E. G. Heinemann, "The relation of apparent brightness to the threshold for difference in luminance," *J. Experiment. Psych.*, vol. 61, pp. 389-399, 1961.

[82] D. Laming, *Sensory Analysis*. New York: Academic Press, 1986.

[83] V. Klymenko, R. E. Johnston, and S. M. Pizer, "Visual threshold as a function of dynamic range and noise in CT images," Presented at *Farwest Image Percept. Conf.*, Tucson, AZ, 1989.

[84] S. M. Pizer, R. E. Johnston, J. B. Zimmerman, and F. H. Chan, "Contrast perception with video displays," *SPIE*, Vol. 318, pp. 223-230, 1982.

[85] S. M. Pizer, J. B. Zimmerman, R. E. Johnston, "Contrast transmission in medical image display," in *Proc. 1st Int. Symp. Med. Imaging Image Interpret. ISMIII*, pp. 2-9, 1982.

[86] R. H. Sherrier and G. A. Johnson, " Regionally adaptive histogram equalization of the chest," *IEEE Trans. Med. Imaging*, vol. MI-6, pp. 1-7, 1987.

[87] J. B. Zimmerman, S. M. Pizer, E. V. Staab, J. R. Perry, W. M. McCartney, and B. C. Brenton, "An evaluation of the effectiveness of adaptive histogram equalization for contrast enhancement," *IEEE Trans. Med. Imaging*, vol. 7, pp. 304-312, 1988.

[88] J. Berkson. "Maximum likelihood and minimum chi-square estimates of the logistic function," *J. Amer. Stat. Ass.*, vol. 50, pp. 130-162, 1955.

[89] W. A. Weibull, "A statistical distribution function of wide applicability." *J. Applied Mechan.*, vol. 18, pp. 292-297, 1951.

[90] D.C. Rogers, R. E. Johnston, and S. M. Pizer, "Effect of ambient light on electronically displayed medical images as measured by luminance-discrimination thresholds," *J. Optical Soc. Amer. A.*, vol. 4, pp. 976-983, 1987.

[91] P. B. Elliot, "Tables of d'," in *Signal Detection and Recognition by Human Observers*, J. A. Swets, Ed., New York: Wiley, pp. 651-684, 1964.

[92] D. Laming, "Precis of Sensory Analysis (and Peer Commentary)," *Behavioral Brain Sci.*, vol. 11, pp. 275-339, 1988.

[93] W. J. McGill and M. C. Teich, "A unique approach to stimulus detection theory in psychophysics based upon the properties of zero-mean gaussian noise (book review of *Sensory Analysis* by D. Laming)," *J. Math Psych.*, vol. 33, pp. 99-108, 1989.

[94] T. Micceri, "The unicorn, the normal curve, and other improbable creatures," *Psycho. Bull.*, vol. 105, no. 1, pp. 156-166, 1989.

[95] D. J. Finney, *Probit Analysis*., Cambridge, England: Cambridge Press, 1971.

[96] J. F. Mackworth, *Vigilance and Habituation*. London, England: Penguin Books, 1970.

[97] D. Kahneman, *Attention and Effort*. Englewood Cliffs, N.J.: Prentice-Hall, 1973.

[98] D. M. Green and R. D. Luce, "Parallel psychometric functions from a set of independent detectors," *Psycho. Rev.*, vol. 82, pp. 483-486, 1975.

[99] B. J. Winer, *Statistical Principles in Experimental Design*. New York: McGraw-HIll, 1971.

[100] T. Corwin, "Efficient threshold curve measurement," Presented at *Psychonomics Soc. Ann. Meet.*, Atlanta. 1989.

[101] A. B. Watson and A. Fitzhugh, "The method of constant stimuli is inefficient," *Percept. Psychophys.*, vol. 47, no. 1, pp. 87-91, 1990.

[102] L. T. Maloney, "Confidence intervals for parameters of psychometric functions," *Percept. Psychophys.*, vol. 47, no. 2, pp. 127-134, 1990.