Department of Nutrition Science Faculty Publications

Department of Nutrition Science

2015

High false positive rates in common sensory threshold tests

Cordelia Running
crunning@purdue.edu

This version of this document is the author's copy.

The final published manuscript can be found at:

https://link.springer.com/article/10.3758/s13414-014-0798-9

Published in *Attention, Perception, and Psychophysics*. 2015, Volume 77: Issue 2, pages 692-700

High false positive rates in common sensory threshold tests

Cordelia A Running

Purdue University, Department of Food Science, West Lafayette, IN

Correspondence to be sent to: Cordelia A Running, Department of Food Science, Purdue University, West Lafayette, IN, USA. E-mail: crunning@purdue.edu

Abstract

Large variability in thresholds to sensory stimuli is frequently observed even in healthy populations. Much of this variability is attributed to genetics and day to day fluctuation in sensitivity. However, false positives are also contributing to variability seen in these tests. In this study, random number generation was used to simulate responses in threshold methods using different "stopping rules": ascending 2-alternative forced choice (AFC) with 5 correct responses; ascending 3-AFC with 3 or 4 correct responses; staircase 2-AFC with 1 incorrect up and 2 incorrect down as well as 1 up 4 down and 5 or 7 reversals; staircase 3-AFC with 1 up 2 down and 5 or 7 reversals. Formulas are presented for rates of false positives in the ascending methods, and curves were generated for the staircase methods. Overall, the staircase methods generally had lower false positive rates, but these methods were even more influenced by number of presentations than ascending methods. Generally, the high rates of error in all these methods should encourage researchers to conduct multiple tests per individual and/or select a method than can correct for false positives, such as fitting a logistic curve to a range of responses.

Key Words: sensory thresholds, type I error, false positive

Introduction

Threshold testing has long been used to evaluate sensory perception in a wide variety of fields (pain research, water contamination, taste sensation, auditory acuity, off flavors, etc). Thresholds are generally grouped into the categories of detection thresholds (lowest concentration of a substance/sensation that is detectable from the background), recognition thresholds (lowest concentration at which a substance/sensation can be identified), and discrimination thresholds (smallest difference in concentration or intensity of a substance/sensation that can be detected in a particular range). Methods have been developed to assess sensory thresholds, all of which require an individual to distinguish the stimulus from a background. Most of these threshold tests are also “forced choice,” meaning that participants are required to make a choice among samples, such as choose a stimulus compared to one or more blanks or choosing a stronger stimulus; if the participant is uncertain which sample to choose, he or she much make a guess. In such cases, participants will occasionally give correct responses accidentally, leading to false positives, or lower than actual thresholds, in the dataset.

In fields of sensory research where participants may be guessing frequently, such as an anosmic person in an olfactory threshold test or when a stimulus is unfamiliar such as in fatty acid “taste” research, rates of false positives in threshold tests become particularly important in interpretation of results. This article is designed to investigate the frequencies of such false positives in sensory threshold experiments, focusing on a few primary techniques common in the field of odor and taste sensitivity research. The high rates of false positives in these methods have been acknowledged (Lawless and Heymann 1998, 2010), but are often not taken into account when analyzing final data. Typical methods for dealing with the false thresholds have been correcting for the proportion of expected “guessers,” which can be done at each concentration step or across the ranges of concentrations, or by fitting psychometric functions to the data which assume a certain rate of false positives. Experiments comparing methods of threshold testing acknowledge that multiple tests, or even multiple methods, will give the most reliable data regarding an individual’s true range of sensitivity, as the variance both among and within subjects in these datasets are high (Boesveldt, de Muinck Keizer, Knol, Wolters, & Berendse, 2009; Doty, McKeown, Lee, & Shaman, 1995; Doty, Smith, McKeown, & Raj, 1994; Haehner et al., 2009; Lotsch, Lange, & Hummel, 2004; Stevens, Cruz, Hoffman, & Patterson, 1995; Tucker & Mattes, 2013). However, comparative data among a variety of testing methods are limited, and most data arise from actual experiments designed to test specific stimuli. While such real world examples of test-retest reliability are extremely valuable, the data from these studies may be less useful in understanding reliability of threshold tests where a stimulus is unfamiliar or even undetectable by certain individuals. These individuals would truly be guessing. The current experiment was designed to observe comparative rates of false positives across a variety of threshold testing methods, using only randomly generated numbers. Thus, the data simulate participants who are guessing. Ideally in sensory threshold testing, participants will eventually reach a concentration at which they can truly identify the stimulus from the blank. The goal of a threshold method would be to isolate these true positive results from the true negative results. However, in a forced choice methodology, false positives will inevitably occur.

The methods emphasized in this article are adaptations of the method of limits: ascending methods (originally from Cain & Rabin, 1989) and "staircase" methods (typically adapted from Deems & Doty, 1987; Doty, Shaman, & Dann, 1984; Wetherill & Levitt, 1965). Within each of these methods, the 2- or 3-alternative forced choice (2-AFC, 3-AFC) tests are common procedures used to determine participant sensitivity at each concentration step. Both were used in the simulation of data. In the 2-AFC paradigm, participants are given 2 samples (one blank, one stimulus) and must identify which contains the stimulus. For the 3-AFC paradigm, participants are given 3 samples (two blanks, one stimulus), and must identify the stimulus. Thus, the 2-AFC method requires some direction (i.e., "Which sample is stronger/sweeter/not water?") while in the 3-AFC method a participant may be instructed simply to identify the "different" sample. Several different "stopping rules" were also investigated in the current analysis, which are discussed in detail in the methods section.

False positives in the ascending method will artificially lower the estimate of a threshold range. In the staircase method, false positives can also contribute to lower estimates, as reversals could occur in the ascending portion of the test prior to the true threshold range being reached. The specific methods analyzed in this article are as follows: 2-AFC ascending method requiring 5 correct identifications, 3-AFC ascending method requiring 3 correct identifications, 3-AFC ascending method requiring 4 correct identifications, 2-AFC staircase method with 1 incorrect up 2 correct down rule, 2-AFC staircase method with 1 incorrect up 4 correct down rule, 3-AFC staircase method with 1 incorrect up 2 correct down rule. The staircase methods were analyzed with both 5 and 7 reversals required to signal the end of the test. Expected rates of false positives for the ASTM method E679, a type of ascending method with a fixed number of stimuli presented to ascertain group threshold values, are also included. The hypotheses were that staircase methods, as the "gold standard" for threshold testing, would exhibit fewer false positives than ascending methods, and that more reversals would lead to fewer false positives.

Methods

Simulated data generation

Excel 2010 was used for generation of random numbers using the formulas RANDBETWEEN(1,2) for 2-AFC or RANDBETWEEN(1,3) for 3-AFC. Two columns of data were generated, the first to represent the actual order of presentation of the stimulus and the second to represent the response of a hypothetical participant. These data mimic what would happen if a participant were guessing, as all positive identifications are due to chance alone. A row of data was counted as a correct identification when the two columns matched. For each row of data, the chance of the "participant" correctly identifying the stimulus is 1/2 for the 2-AFC and 1/3 for the 3-AFC paradigms.

Ascending method of limits

In the ascending method of limits, the test begins at a low concentration of the stimulus and the concentration is increased until the participant can identify the stimulus correctly. The samples are presented in random order. The participant selects the sample they believe contains the stimulus, and the test is repeated based on the participant's response. If the participant is correct, the same concentration of stimulus is presented in the next round. If the participant is incorrect, the next higher concentration of stimulus is presented. This continues until the participant can reliably identify the stimulus according to a predetermined "stopping rule," or until all sample concentrations have been tested. The threshold in this test may either be the actual concentration at which the stopping criterion was met, or the mean of that concentration and the concentration below (calculated either as the mean of the log concentration or the geometric mean, see Lawless, 2013).

For the current analysis, the ascending method of limits was analyzed in three ways. Using the 2-AFC paradigm, 5 sequential correct responses were required. Using the 3-AFC paradigm, analysis was conducted on both 3 sequential correct responses and 4 sequential correct responses. Formulas were derived for the expected rate of false positives for each method and matched to simulated data curves, in order to confirm the accuracy of the formulas. For data simulation, fifty rows of data were generated for each method, each row of data representing one presentation of samples to a participant. If the stopping criterion was met (3, 4, or 5 "correct" responses), the row number at which the stop occurred was noted (i.e., the "run length" of the test). The data were refreshed 100 times to simulate data from 100 participants.

Staircase method of limits

In the staircase method of limits, the test begins ideally in the center of the expected range of threshold concentrations. Participants are presented with blank and stimulus samples in random order as before according to the 2- or 3-AFC paradigm. If a participant's response is incorrect, then the trial is repeated with the next higher concentration of stimulus (the "1 up" rule). If the participant is correct, then next trial is typically repeated at the same concentration. For the "2 down" rule, if the participant is correct at again at the same concentration, then the next trial is conducted with the lower concentration of stimulus. For the "4 down" rule, the participant must be correct at the same concentration 4 times sequentially before the concentration is lowered. An example of this method for a "1 up 2 down" rule is given in Figure 1. For the simulated data, the "1 up 2 down" rule was employed with both the 2-AFC and 3-AFC paradigms, and the "1 up 4 down" rule was employed with the 2-AFC paradigm. The staircase method continues until a predetermined number of "reversals" occur, i.e. switching from correct identification to incorrect identification. In the simulated data, analysis was conducted with both 5 reversals and 7 reversals.

Data were generated as before. For the "1 up 2 down" rule, a pattern of one incorrect response followed by 2 correct responses (ICC) or two correct responses followed by one incorrect response (CCI) indicates a reversal. The first ICC or CCI is one reversal, and each subsequent ICC or CCI is two reversals (see Figure 1). Thus, for 5 reversals, three ICC or CCI patterns are needed to complete the task, while for 7 reversals four of these patterns are needed. For the "1 up 4 down" rule, the pattern ICCCC or CCCCI indicates reversals, still with 3 or 4 repeats required to observe 5 or 7 reversals, respectively. A column in Excel was generated to indicate whether the response was correct or incorrect, and the number of ICC(CC) or CC(CC)I patterns was counted over 50 (for 1 up 2 down) or 100 (for 1 up 4 down) rows of data, to simulate 50 or 100 presentations of sample (the greater number of presentations was generated for the 1 up 4 down rule because of the larger number of presentations required in this test). Such long run lengths are not typical of most sensory threshold tests, especially in gustation and olfaction, but were used to observe the asymptotes and changes in the curves over time. The data were refreshed 100 times to represent 100 participants, and the rows at which correct numbers of reversals was reached was recorded. This was done for all versions of the staircase method. As formulas for predicting the expected rate of false positives for staircase methods would be very complex, and as attempts to fit logistic regression curves to the data yielded poor fit in the lower ranges of run length, data were again refreshed 500 times for each of the staircase methods and Excel was used to generate smoothed curves based on these large datasets. These values were used to determine at what run lengths the methods would be expected to exceed 5% and 10% of the participants giving false thresholds (assuming all participants are guessing), as these are typical α levels.

ASTM International E679 – 04

ASTM standard E679 – 04 is designed for small datasets (less than 100 presentations) to estimate group, not individual, thresholds (ASTM 2011). The method is based on the concept that thresholds are probability functions, where at low concentrations the probability of an individual detecting the stimulus is zero and at high concentrations the probability is 1 (corrected for guessing). Samples are prepared in 5-8 concentration steps, each differing by a factor of 2 to 4 (e.g., for a factor of 3: x/27, x/9, x/3, x, 3x, 9x, 27x). Thresholds of each individual are calculated as the geometric mean (or mean of the logarithm of the concentrations) of the last incorrect response and the first correct response, after which no other incorrect responses were given ("last reversal"). Group means for thresholds are the geometric mean (or mean of the logarithm of the concentrations) of all participant mean thresholds. In the current data, expected false positives were calculated for each concentration step. Data were not simulated for this method, as the rates of expected false positives at each presentation are easily calculable.

Table 1 gives a summary of the methods and stopping rules tested in the simulated data. Additionally, this table lists the minimum number of presentations (i.e., shortest run length) required in order for a participant to complete the test. For example, in the ascending method, to achieve 4 correct identifications, at least 4 presentations are required. In the staircase method with a 1 up 4 down rule, 15 presentations are required at minimum to achieve 5 reversals.

Results

ASTM E679

Equations used to calculate expected false positives at each of 7 concentration steps are shown in Table 2, along with the calculated rates. Note that in order for the criterion of the "last reversal" rule to be met, an incorrect response must precede the correct responses for steps 2-7, hence the 2/3 factor in the formula. Rates of false positives are lower, as expected, for the lower concentration steps and increase with the higher concentration steps. This is clearly a function of fewer correct responses required to achieve a false positive at the higher concentrations.

Ascending methods of limits

Figure 2 shows the cumulative rate of false positives in the 5ASC, 3ASC, and 4ASC method of limits over the first 50 presentations (run length) using the formulas given in Table 3. While 50 presentations would be an uncommonly high run length for a gustatory or olfactory threshold test, this run length is shown to observe how the rates of false positives begin to asymptote with more presentations. The simulated data curved fit very well with the formula generated curves, thus these data are not shown. The 3ASC (3-AFC with 3 correct responses) displayed the highest rates of false positives, followed by the 5ASC (2-AFC with 5 correct responses) then the 4ASC (3-AFC with 4 correct responses).

Staircase method of limits

Figure 3a shows the cumulative rate of false positives for the staircase methods. Figure 3a shows the methods with 500 simulated participants, and Figure 3b shows these methods shifted for the minimum required run length in order to complete the test (from Table 1). The 2-12-5 and -7REV (2-AFC, 1 up 2 down with 5 or 7 reversals) showed very rapid increases of false positives with run length. Slower increases in error were observed for the 3-12-5 and -7REV (3-AFC versions) methods. The 2-14-5 and -7REV methods (2-AFC with 1 up 4 down) showed the lowest rates of error of any tests; however, these two versions of the staircase methods require more presentations (longer run length) due to the larger number of trials needed before it's even possible to meet the stopping criteria. Again, the run lengths of 100 presentations are not reasonable for olfactory or gustatory tests, but are included to observe the asymptotes of the curves and to be able to compare the different methods to each other.

Comparison of false positives in various tests

Table 4 shows where each method, using the generated formulas for the ascending methods and the large datasets for the staircase methods, crosses 5% and 10% rates. The table also shows this analysis shifted to account for the minimum number of presentations required to complete the task. Figure 4 shows comparisons of all methods of limits, (A) 2-AFC paradigms and (B) 3-AFC paradigms, shifted to account for the minimum run length required to complete the test. For the 2-AFC paradigm, the staircase method with a 1 up 4 down clearly results in much lower error than any of the other methods. For the 3-AFC paradigm, the staircase methods may be preferable if run lengths can be kept short, under a total of about 18 presentations (9 required to complete the test, crosses over 4ASC method at 9 in the figure) for 5 reversals and under 31 presentations (12 to complete the test, crosses 4ASC method at about 18 in the figure) for 7 reversals. As seen in figure 4, the slope of rate of guessing increases with run length for staircase methods, while the slope decreases for ascending methods.

Discussion

The high rates of false thresholds observed in the current data would increase variability in sensory threshold studies both within and between subjects, but only when participants are guessing. This variability is clearly dependent on the method and stopping rule used in the test as well as upon the method for data analysis. The impact of the variability and type of test, as well as some proposed methods to deal with the rates of false stops, are discussed below.

The data presented here show the stricter stopping rules result in lower rates of false stops, as should be expected. Staircase methods have lower rates of error when the run lengths are minimized, but very rapidly increase in false stops as the number of presentations increases. Notably, the longer run lengths will also contribute to fatigue on the part of the participant, especially in experiments on olfaction and gustation. Thus, for longer run lengths, staircase methods become less reliable than ascending methods. The staircase method, particularly the 3-AFC paradigm with 7 reversals, has been considered a “gold standard” of sensory threshold testing, particularly for olfaction (Lotsch et al., 2004), and experiments comparing ascending to staircase methods generally report that staircase methods are more reliable and show less variability (Doty et al., 1995; Linschoten, Harvey, Eller, & Jafek, 2001; Tucker & Mattes, 2013). However, the data presented here indicate caution should be used with the staircase methods, and attempts should be made to minimize the run length of the test not just for the sake of limiting participant fatigue, but also for the sake of fewer artificially low thresholds. Given the high slopes of the staircase methods as the number of presentations increases, the 4ASC method could be a viable alternative for some experimental settings.

The reliability of human sensory threshold tests for olfaction and gustation is often low (Doty et al., 1995; Lawless, Thomas, & Johnston, 1995; Stevens et al., 1995; Stevens & Dadarwala, 1993). While some studies indicate test-retest correlation coefficients of staircase methods for olfactory thresholds above 0.8 (Lotsch et al., 2004, Doty et al. 1995, Haehner et al. 2009), others demonstrate coefficients in the range of 0.6-0.7, with even lower correlations over longer periods of time (Linschoten et al. 2001). Taste thresholds often show test-retest coefficients around 0.6 or less (McMahon et al. 2001, Stevens et al. 1995, Linschoten et al. 2001). Large variability has also been observed within subjects even in the short term for these chemosensory systems (Jaeger, de Silva, & Lawless, 2014; McMahon, Shikata, & Breslin, 2001; Stevens, Cain, & Burke, 1988). Much of this variability is due to the type of test employed, the sensory modality being tested, as well as physiological or psychological effects within a person, as all threshold tests require careful attention to detail and the ability to make fine distinctions. Additionally, factors such as familiarity with a stimulus, learning (Lawless & Heymann 1998, 2010; ASTM 2011, Tucker & Mattes 2013), dilution step sizes, and level of feedback on whether or not a response is correct (Doty et al. 2003) can also influence test-retest reliability. However, current data indicate that a large amount of variability may also be attributable to the tests themselves, as higher rates of false positives may occur than previously assumed. Further, previous studies have observed that more stringent stopping rules tend to yield higher thresholds (Peng, Jaeger, & Hautus, 2012), which would be in agreement with the rates of false positives observed in the current data.

For the ascending method, the stopping rules have typically been set by the number of presentations needed to below a type I error of 5%; i.e., a 2-AFC paradigm may require 5 correct responses because the probability is $(1/2)^5 = 3.1\%$ and a 3-AFC paradigm may require 3 correct responses as $(1/3)^3 = 3.7\%$. As originally noted by Lawless and Heymann (1998), this approach does not account for multiple testing, which is why observed rate of guessing correctly in the simulated data is much higher than given by the stopping rule alone. The longer the test continues (longer run length, more presentations), the more likely a false positive will occur because there are more opportunities for the event to occur. The concept is the same as with lottery tickets: it is very unlikely that "you" will win the lottery, but it is very likely that "someone" will win the lottery.

False positives in threshold tests can only occur when a participant is guessing. Because of this, a false positive must fall below that the range of concentrations of participant's actual threshold range. In ascending methods, the true threshold range may not be reached at all, and underestimates could be quite large. In staircase methods, false positives would create reversals below the true threshold range, again contributing to underestimation and also potentially prolonging the test and providing more opportunities for additional false positives. If the concentration is above the threshold region, the participant should not be guessing so the response will not contribute to false positives, unless fatigue or adaptation are interfering with determinations. Thus, beginning the test as close as possible to the true range of a participant's threshold will reduce the opportunity for false positives in the responses. For staircase methods, the test should ideally begin at the hypothesized threshold region for that individual, and for the ascending method, the test should begin just below the threshold. This will reduce the run length of the test. Reliability has already been correlated with the run length of threshold tests (Doty et al., 1995). Data in the current analysis show that this is not only due to decreased fatigue for the participant, but also to fewer opportunities for false positives. Reports, and data from the author's current laboratory, typically give run lengths ranging from 10-25, with ascending methods generally giving shorter run lengths than staircase methods (Linschoten et al., 2001; Stevens et al., 1995). Thus, researchers may want to analyze average run lengths in an experiment before finalizing results.

Starting the threshold test near an individual's threshold region means that different individuals will begin the test at different concentrations. This would require some knowledge of the individuals' sensitivities, again requiring at least two tests per person: one to give an initial idea of the threshold, and the second to test the accuracy of that threshold. Numerous studies have already reported that multiple thresholds tests are required to give reliable assessments of an individual's sensitivity to a particular compound (McMahon et al., 2001; Stevens et al., 1995; Stevens & Dadarwala, 1993; Tucker & Mattes, 2013). Typically this has been attributed to natural variation in a subjects' ability to detect the compound or to learning effects with multiple tests. However, the data in the current study indicate that much of this variability, leading to the need for multiple tests to assess a single individual, may also be due to false positives. While a range of sensitivity should still be expected, the breadth of this range will be expanded if artificially low estimates are included in the data. Reducing the rates of false positives could potentially decrease the number of tests needed to assess not only the overall sensitivity of a subject to a sensation, but also could give a clearer picture of the true range of an individual's day to day sensitivity. For a fast assessment, a brief ascending series of stimuli could be presented (for example, 5 concentrations each ½ or a full logarithmic dilution apart, depending on the stimulus and prior knowledge of differences in sensitivity among individuals), and the responses to that series of presentations could be used to guide a second test with a finer set of dilutions (the more common ¼ logarithmic dilution apart). In staircase methods, such differences in step sizes may be built into the procedure, beginning with larger step sizes and reducing the step size in the perithreshold region after observing at least one reversal. This also reduces the number of presentations in the procedure. For studies with novel stimuli on which prior data are unavailable, multiple testing visits would be needed to first assess the range of sensitivity across subjects and then accurately assess the individual subjects' sensitivity range.

For situations in which multiple tests visits are impractical, a method should be used that corrects for guessing. The common technique for this is to fit a logistic curve to the rates of correct/incorrect responses over a range of concentrations. Techniques for adapting the ASTM E679 (Lawless, 2010) or general ascending methods (Hough, Methven, & Lawless, 2013) to correct for guessing have already been proposed. These two proposed modifications basically correct participant's data by taking into account their subsequent responses, higher in the concentration series, and other participant's performance at each concentration. Modifying these methods to correct for guessing, as well as for participants whose sensitivity falls outside the range of tested concentrations, allows for a faster collection of a larger amount of data than testing individuals multiple times. However, these techniques may be less useful for assessing an individual's sensitivity accurately. While the techniques have been used to find differences between groups (Hough et al., 2013), using the technique to assess an individual in a clinical setting may be more difficult.

Another suggestion for improving the quality of data while minimizing run length is to alter the application of the stopping rule in the ascending method. Typically, if a response is correct, the same concentration of stimulus is presented until the participant is correct the predetermined number of times. However, in order to reduce the number of presentations, the same concentration could be presented 2 or 3 times, then the next higher concentration could be presented. The stopping rule of 4 or 5 correct responses could still be used, but the correct responses would be spread across numerous different concentrations. Then, if a participant gives an incorrect response, the test would continue with fewer overall presentations. For example: At concentration 6, the participant is correct 3 times. Instead of giving concentration 6 again, concentration 5 (more concentrated) is given. If the participant is correct at concentration 5, a stopping rule of "4 correct" would be met. If they are incorrect, the test could continue, with fewer overall presentations than would have been used if the participant had been tested 4 times at concentration 6, and given an incorrect response on the 4th presentation. Indeed, if a participant's true threshold were at concentration 6, then that individual should even more easily detect the stimulus at concentration 5.

Again, it should be noted that false positives in sensory threshold tests are only a problem when participants are guessing. Generally, by testing many participants, or by testing participants multiple times, the overall effect of these false positives on conclusions and observations may be small. However, the high rates of false positives should be particularly concerning when the research concerns novel or poorly defined sensory stimuli. For instance, false positives should be a concern in the field of non-esterified fatty acid (NEFA) "taste" research. Most of the work conducted in this field has focused on taste thresholds for NEFA, and whether such thresholds correlate to other dietary or physical attributes or habits of humans (for reviews, see Passilly-Degrace et al., 2014 and Running, Mattes, & Tucker, 2013). While data indicate there are mechanisms in humans to perceive these compounds as a "taste," human participants in the studies may be guessing frequently during the threshold tests, as published data indicate very large ranges of sensitivity to these compounds (Running & Mattes, 2014; Running, Mattes, & Tucker, 2013; Tucker, Edlinger, & Mattes 2014; Tucker & Mattes 2013). With such a large range of potentially detectable concentrations, starting the test near the hypothesized threshold is difficult, and the required longer run length of the test will thus increase the chance of false positives. Work with repeated testing indicates that some participants improve (lower their thresholds) over time (Tucker, Edlinger, & Mattes 2014; Tucker & Mattes 2013). Such learning effects are to be expected in threshold testing (ASTM 2011; Lawless & Heymann 1998, 2010), but particularly of interest is the observation some participants continued to improve over all 10 visits for the ascending method while in the staircase method the maximum learning effect was observed by visit 7 (Tucker & Mattes 2013). Potentially, this could be an effect of false positives on the mean threshold value. In the ascending method, participants began below their previously measured threshold while in the staircase method participants always began at the same concentration step. Thus, every time a false stop occurred in the ascending method, that participant would begin the test even further away from his or her true threshold region on the next visit, and would thus increase the run length of the test before that true threshold range could be reached. This would increase the likelihood of a false stop on this next visit. Consequently, basing each study visit's starting concentration on the previous visit's threshold may not be ideal when conducting multiple tests with the ascending method. At very least, the participant's ability to detect the lower concentrations should be verified with a more stringent test if large improvements are continually observed in multiple ascending tests.

Conclusions

Rates of false positives in threshold tests were much higher than would have been predicted by analyzing stopping rules alone. The data generated by random numbers agreed with previous observations, that longer run lengths (more presentations) will increase the variability in the tests, and that staircase methods may be more reliable than ascending methods. However, it should be noted, as observed in the figures, that for staircase methods rates of false positives increase very rapidly with the increasing run length of the test. In some circumstances the ascending methods may be preferable to reduce the total number of presentations and thus the chance of guessing correctly. Generally, applying a method that can correct for the chance of guessing is preferable to avoid the high rates of artificially low thresholds observed in these data, and multiple tests per participant may allow for observation of when a false threshold occurs.

Acknowledgements

Special thanks are due to Dr. Richard Mattes and Dr. Bruce Craig for discussions on the methods and data presented.

Table 1: Methods and stopping rules tested

Method	Choices	Stopping rule		Abbreviation	Minimum Run Length
Ascending	2-AFC	5 sequential correct		5ASC	5
	3-AFC	3 sequential correct		3ASC	3
		4 sequential correct		4ASC	4
Staircase	2-AFC	1 up 2 down	5 reversals	2-12-5REV	9
			7 reversals	2-12-7REV	12
		1 up 4 down	5 reversals	2-14-5REV	15
			7 reversals	2-14-7REV	20
	3-AFC	1 up 2 down	5 reversals	3-12-5REV	9
			7 reversals	3-12-7REV	12
ASTM E679	3-AFC	Last reversal from incorrect to correct		E679	7 (fixed)

Table 2: Calculations for ASTM E679

Probability of a false positive at step 1 (most dilute)	$\frac{1^7}{3}$	Step 1: 0.0%
Probability of a false positive at step 2-7 (where *i* is the step number, and step 7 is the most concentrated)	$\left(\frac{1}{3}\right)^{8-i} \times \frac{2}{3}$	Step 2: 0.1% Step 3: 0.3% Step 4: 0.8% Step 5: 2.5% Step 6: 7.4% Step 7: 22.2%

Table 3: Ascending methods false positive rate by run length

5ASC	
Run length (i)	Probability of stopping at i [$P(i)$]
5	$\frac{2^{i-5}}{2^i}$
6-10	$\frac{2^{i-5}-2^{i-6}}{2^i}$
11-15	$\frac{(2^{i-5}-2^{i-6})-(2^{i-10}-2^{i-11})}{2^i}$
16-20	$\frac{(2^{i-5}-2^{i-6})-(2^{i-10}-2^{i-11})-(2^{i-15}-2^{i-16})}{2^i}$
Etc.	
Cumulative probability of stopping at or before i: $1-\{[1-P(i)]\times[1-P(i-1)]\times[1-P(i-2)]\times\cdots\times[1-P(i-a)]\}$ Where $a=i-5$	
3ASC	
Run length (i)	Probability of stopping at i [$P(i)$]
3	$\frac{3^{i-3}}{3^i}$
4-6	$\frac{3^{i-3}-3^{i-4}}{3^i}$
7-9	$\frac{(3^{i-3}-3^{i-4})-(3^{i-6}-3^{i-7})}{3^i}$
10-12	$\frac{(3^{i-3}-3^{i-4})-(3^{i-6}-3^{i-7})-(3^{i-9}-3^{i-10})}{3^i}$
Etc.	
Cumulative probability of stopping at or before i: $1-\{[1-P(i)]\times[1-P(i-1)]\times[1-P(i-2)]\times\cdots\times[1-P(i-a)]\}$ Where $a=i-3$	
4ASC	
Run length (i)	Probability of stopping at i [$P(i)$]
4	$\frac{3^{i-4}}{3^i}$
5-8	$\frac{3^{i-4}-3^{i-5}}{3^i}$
9-12	$\frac{(3^{i-4}-3^{i-5})-(3^{i-8}-3^{i-9})}{3^i}$
13-16	$\frac{(3^{i-4}-3^{i-5})-(3^{i-8}-3^{i-9})-(3^{i-12}-3^{i-13})}{3^i}$
Etc.	
Cumulative probability of stopping at or before i: $1-\{[1-P(i)]\times[1-P(i-1)]\times[1-P(i-2)]\times\cdots\times[1-P(i-a)]\}$ Where $a=i-4$	

Table 4: Run lengths that exceed 5% or 10% type I error

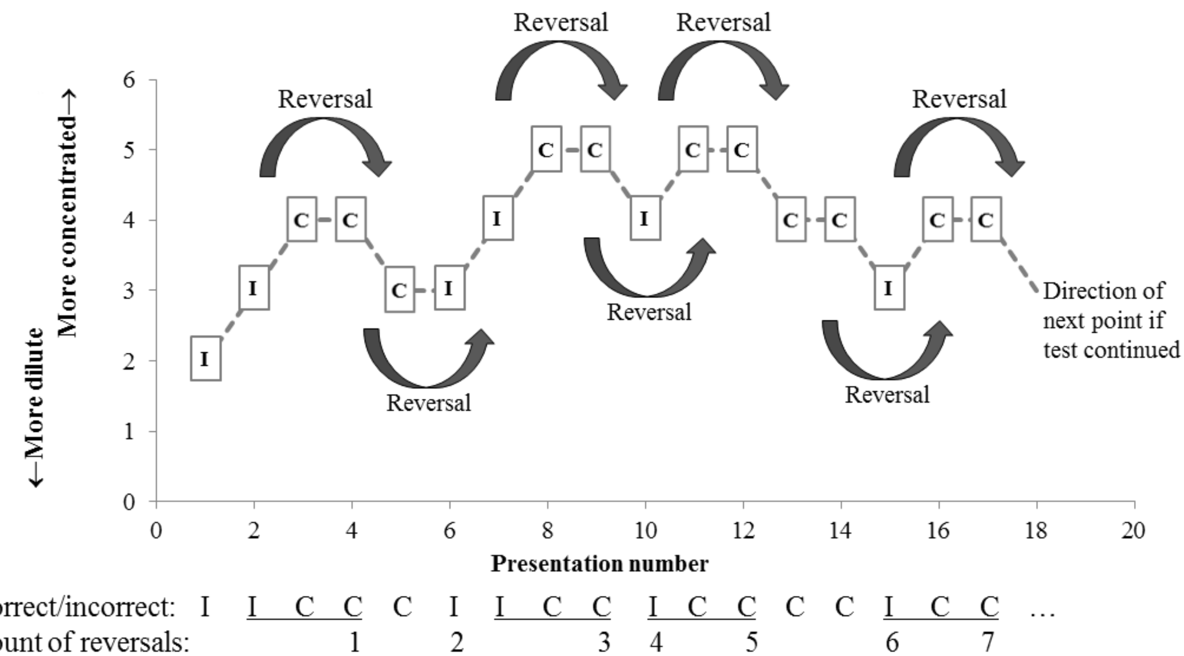
	Run length when exceeds:		Run length past minimum when exceeds:	
Method	5%	10%	5%	10%
5ASC	7	10	2	5
3ASC	4	4	1	2
4ASC	9	16	5	12
3-12-5REV	17	19	8	10
3-12-7REV	23	28	11	16
2-12-5REV	12	13	3	4
2-12-7REV	18	20	6	8
2-14-5REV	34	44	19	29
2-14-7REV	54	70	34	50

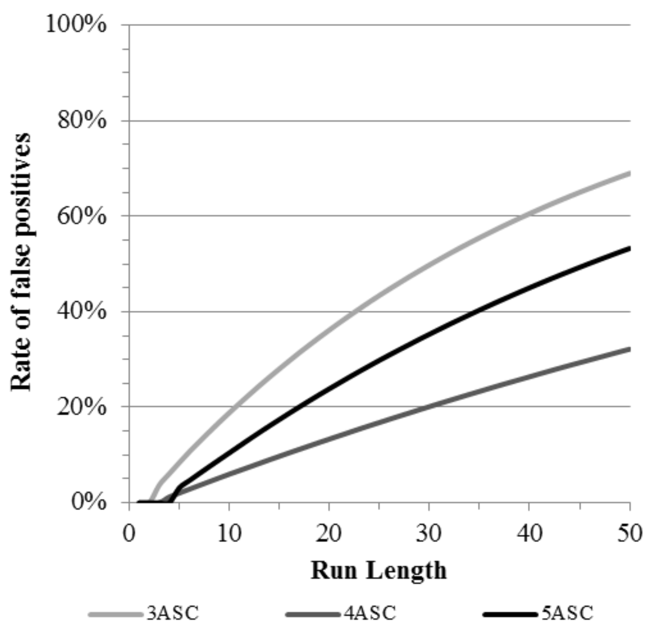
Figure 1: Illustration of staircase method and patterns of correct/incorrect responses for reversals

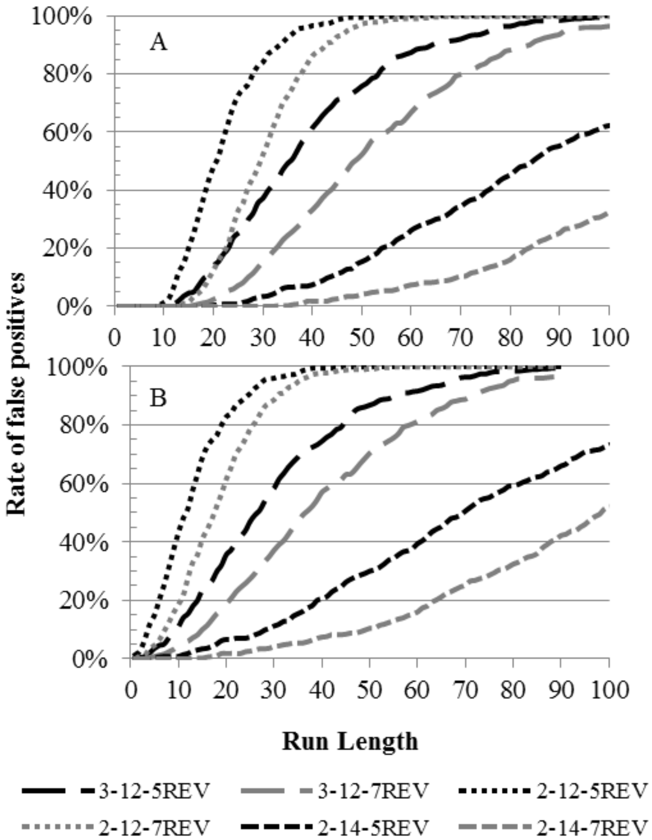
Figure 2: False positive rates by run length for ascending method 2-AFC with 5 correct responses (5ASC) and 3-AFC with 3 (3ASC) or 4 (4ASC) correct responses required as stopping rule.

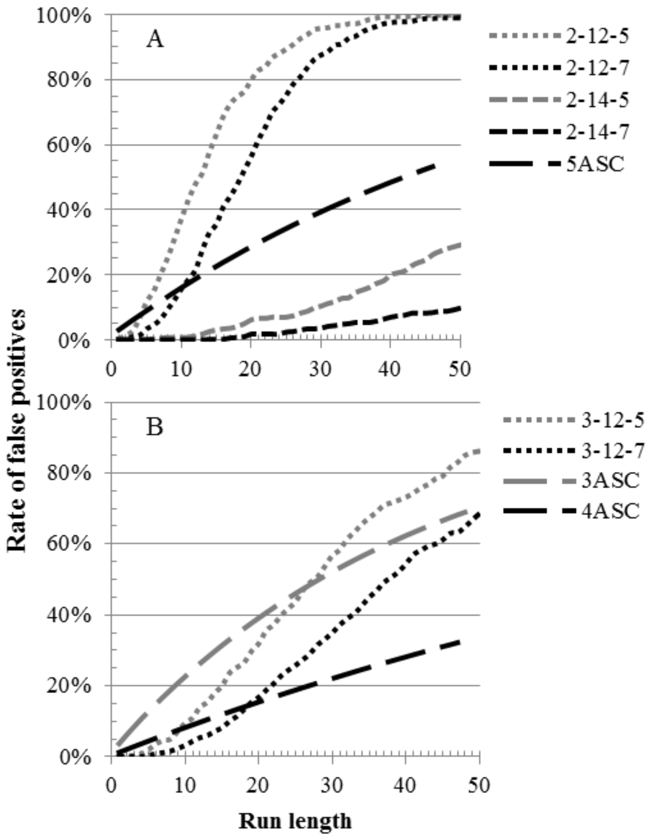
Figure 3: False positive rates by total run length (A) or run length shifted for minimum required to achieve stopping rule (B) for staircase methods. 3-12-5REV: 3AFC method 1 up 2 down rule and 5 reversals, 3-12-7REV: 3AFC method 1 up 2 down rule and 7 reversals, 2-12-5REV: 2AFC method 1 up 2 down rule and 5 reversals, 2-12-7REV: 2AFC method 1 up 2 down rule and 7 reversals, 2-14-5REV: 2AFC method 1 up 4 down rule and 5 reversals, 2-14-7REV: 2AFC method 1 up 4 down rule and 7 reversals.

Figure 4: Comparison of 2-AFC (top) and 3-AFC (bottom) staircase and ascending methods, using run length shifted for minimum required to achieve stopping rule. 3-12-5REV: 3AFC method 1 up 2 down rule and 5 reversals, 3-12-7REV: 3AFC method 1 up 2 down rule and 7 reversals, 2-12-5REV: 2AFC method 1 up 2 down rule and 5 reversals, 2-12-7REV: 2AFC method 1 up 2 down rule and 7 reversals, 2-14-5REV: 2AFC method 1 up 4 down rule and 5 reversals, 2-14-7REV: 2AFC method 1 up 4 down rule and 7 reversals, 5ASC: 2AFC method with 5 correct responses, 3ASC: 3AFC method with 3 correct responses, 4ASC: 3AFC method with 4 correct responses.


Reversal
Reversal
Reversal
Reversal
Reversal
Reversal
Reversal
←More dilute
More concentrated→
Presentation number
Direction of next point if test continued
Correct/incorrect: I I C C C I I C C I C C C C I C C …
Count of reversals: 1 2 3 4 5 6 7


Rate of false positives
Run Length
3ASC
4ASC
5ASC


A
B
Rate of false positives
Run Length
0%
20%
40%
60%
80%
100%
0
10
20
30
40
50
60
70
80
90
100
3-12-5REV
3-12-7REV
2-12-5REV
2-12-7REV
2-14-5REV
2-14-7REV



A
B
2-12-5
2-12-7
2-14-5
2-14-7
5ASC
3-12-5
3-12-7
3ASC
4ASC
Rate of false positives
Run length
0%
20%
40%
60%
80%
100%
0
10
20
30
40
50

References

ASTM. Standard E679-04 (2011), "Standard Practice for Determination of Odor and Taste Thresholds By a Forced-Choice Ascending Concentration Series Method of Limits".

Boesveldt, S., de Muinck Keizer, R. J., Knol, D. L., Wolters, E., & Berendse, H. W. (2009). Extended testing across, not within, tasks raises diagnostic accuracy of smell testing in Parkinson's disease. *Mov Disord, 24*(1), 85-90. [

Cain, W. S., & Rabin, M. D. (1989). Comparability of 2 Tests of Olfactory Functioning. *Chem Sens, 14*(4), 479-485.

Deems, D. A., & Doty, R. L. (1987). Age-related changes in the phenyl ethyl alcohol odor detection threshold. *Trans Pa Acad Ophthalmol Otolaryngol, 39*(1), 646-650.

Doty, R. L., Diez, J. M., Turnacioglu, S., McKeown, D. A., Gledhill, J., Armstrong, K., & Lee, W. W. (2003). Influences of feedback and ascending and descending trial presentations on perithreshold odor detection performance. *Chem Senses, 28*(6), 523-526.

Doty, R. L., McKeown, D. A., Lee, W. W., & Shaman, P. (1995). A study of the test-retest reliability of ten olfactory tests. *Chem Senses, 20*(6), 645-656.

Doty, R. L., Shaman, P., & Dann, M. (1984). Development of the University of Pennsylvania Smell Identification Test: a standardized microencapsulated test of olfactory function. *Physiol Behav, 32*(3), 489-502.

Doty, R. L., Smith, R., McKeown, D. A., & Raj, J. (1994). Tests of human olfactory function: principal components analysis suggests that most measure a common source of variance. *Percept Psychophys, 56*(6), 701-707.

Haehner, A., Mayer, A. M., Landis, B. N., Pournaras, I., Lill, K., Gudziol, V., & Hummel, T. (2009). High test-retest reliability of the extended version of the "Sniffin' Sticks" test. *Chem Senses, 34*(8), 705-711.

Hough, G., Methven, L., & Lawless, H. T. (2013). Survival Analysis Statistics Applied to Threshold Data Obtained from the Ascending Forced-Choice Method of Limits. *J Sens Stud, 28*(5), 414-421.

Jaeger, S. R., de Silva, H. N., & Lawless, H. T. (2014). Detection Thresholds of 10 Odor-active Compounds Naturally Occurring in Food Using a Replicated Forced-Choice Ascending Method of Limits. *J Sens Stud, 29*(1), 43-55.

Lawless, H. T. (2010). A simple alternative analysis for threshold data determined by ascending forced-choice methods of limits. *J Sens Stud, 25*(3), 332-346.

Lawless, H. T. (2013). Psychophysics I: Introduction and Thresholds *Quantitative Sensory Analysis* (pp. 1-23). Chichester, UK: John Wiley & Sons.

Lawless, H. T., & Heymann, H. (1998). *Sensory Evaluation of Food: Principles and Practices* (1st ed.). New York, NY: Chapman & Hall.

Lawless, H. T., & Heymann, H. (2010). *Sensory Evaluation of Food: Principles and Practices* (2nd ed.). New York, NY: Springer.

Lawless, H. T., Thomas, C. J., & Johnston, M. (1995). Variation in odor thresholds for l-carvone and cineole and correlations with suprathreshold intensity ratings. *Chem Senses, 20*(1), 9-17.

Linschoten, M. R., Harvey, L. O., Jr., Eller, P. M., & Jafek, B. W. (2001). Fast and accurate measurement of taste and smell thresholds using a maximum-likelihood adaptive staircase procedure. *Percept Psychophys, 63*(8), 1330-1347.

Lotsch, J., Lange, C., & Hummel, T. (2004). A simple and reliable method for clinical assessment of odor thresholds. *Chem Senses, 29*(4), 311-317.

McMahon, D. B., Shikata, H., & Breslin, P. A. (2001). Are human taste thresholds similar on the right and left sides of the tongue? *Chem Senses, 26*(7), 875-883.

Passilly-Degrace, P., Chevrot, M., Bernard, A., Ancel, D., Martin, C., & Besnard, P. (2014). Is the taste of fat regulated? Biochimie, 96, 3-7.

Peng, M., Jaeger, S. R., & Hautus, M. J. (2012). Determining odour detection thresholds: Incorporating a method-independent definition into the implementation of ASTM E679. *Food Qual Prefer*, *25*(2), 95-104.

Running, C. A., & Mattes, R. D. (2014). Different oral sensitivities to and sensations of short, medium, and long chain fatty acids in humans. *Am J Physiol Gastrointest Liver Physiol, 307*, G381-389.

Running, C. A., Mattes, R. D., & Tucker, R. M. (2013). Fat taste in humans: Sources of within- and between-subject variability. *Prog Lipid Res, 52*(4), 438-445. doi: 10.1016/j.plipres.2013.04.007

Stevens, J. C., Cain, W. S., & Burke, R. J. (1988). Variability of Olfactory Thresholds. *Chem Sens, 13*(4), 643-653.

Stevens, J. C., Cruz, L. A., Hoffman, J. M., & Patterson, M. Q. (1995). Taste sensitivity and aging: high incidence of decline revealed by repeated threshold measures. *Chem Senses*, *20*(4), 451-459.

Stevens, J. C., & Dadarwala, A. D. (1993). Variability of olfactory threshold and its role in assessment of aging. *Percept Psychophys, 54*(3), 296-302.

Tucker, R. M., Edlinger, C., Craig, B. A., & Mattes, R. D. (2014). Associations between BMI and fat taste sensitivity in humans. Chem Sens, 39(4), 349-357.

Tucker, R. M., & Mattes, R. D. (2013). Influences of Repeated Testing on Nonesterified Fatty Acid Taste. *Chem Sens, 38*(4), 325-332.

Wetherill, G. B., & Levitt, H. (1965). Sequential Estimation of Points on a Psychometric Function. *Brit J Math Stat Psy, 18*(1), 1-10.