The Effect of F0 Rate on Accent Perception in English
Author(s): Kazue Hata and Yoko Hasegawa
*Proceedings of the Seventeenth Annual Meeting of the Berkeley Linguistics Society: General Session and Parasession on The Grammar of Event Structure* (1991), pp. 121-129

# The Effect of F0 Fall Rate on Accent Perception in English[1]

Kazue Hata
Speech Technology Laboratory

Yoko Hasegawa
University of California, Berkeley

## 1. Introduction

Onishi (1942) points out that the function of accent is to differentiate the meaning, or to make prominent a portion, of words or phrases, and that accent is an impressionistic sum of any features that could serve these purposes. It has been widely recognized that in English four psychoacoustic dimensions influence the perception of accent (stress): pitch, length, loudness, and sound quality. In neutral declarative intonation, the accented syllables carry, relative to non-accented syllables, higher fundamental frequency (F0), longer duration, higher amplitude, and such different spectral patterns as in energy distribution among vowel formants.

Fry (1958) conducted perceptual experiments with synthetic noun-verb pairs in which the distinction is made by the accent placement, e.g. *súbject* vs. *subjéct*. He found that the increase in vowel duration of the second syllable can cause a perceived accent shift from noun *súbject* to verb *subjéct*. The increase in amplitude has a similar effect, although to a lesser magnitude. As for the ranking between F0 and duration cues, typically the former outweighs the latter. Therefore, in Fry's experiment, the most significant cue to the accent was F0, followed by duration and then by amplitude.

Naturally, then, one may think that accent location is determined by the location of F0 peak. However, this is not always the case. The perceived accent and the actual F0 peak sometimes do not synchronize without listeners detecting this desynchronization (Lehiste and Peterson 1961, Neustupný 1966, Sugito 1972, Hasegawa and Hata 1988, Hata and Hasegawa 1988). In Japanese, for example, the listener perceives an accent on a syllable even when the F0 peak does not occur on it. In Hasegawa and Hata (1988), we presented the following pair from our production data. They are part of the word *namida* 'tear', in which the lexical accent falls on the first syllable. In the figure on the right, the F0 peak is clearly on /i/, and yet the word was perceived as /námida/.

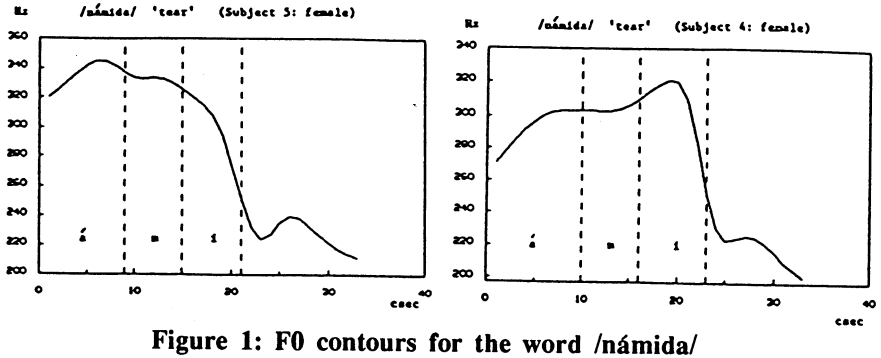


**Figure 1: F0 contours for the word /námida/**

Sugito (1972) found that this illusory accent is due to the F0 contour falling after the peak: if the peak is followed by a steep F0 fall, the listener perceives an accent on the preceding syllable, as shown in Figure 2.[2]

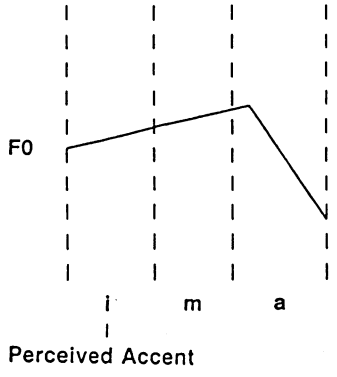


**Figure 2: Perceived accent and the actual F0 peak**

This phenomenon of illusory accent explains why the native Japanese listener perceives an accent on a devoiced vowel. Even though high F0 does not occur on the devoiced vowel, the F0 fall on the following syllable forces the listener to associate an accent with that vowel.

Hata and Hasegawa (1988) found that there is a positive correlation between the F0 peak location and the F0 fall rate immediately after the peak in those utterances where the perceived accent was shifted from the location signaled by the F0 peak. The later the F0 peak occurred in the second syllable, relative to the syllable boundary, the greater the fall rate necessary for the listener to associate the accent with the first syllable. For example, when the F0 peak was at about 50% of the second syllable, the majority of the subjects judged the first syllable to be accented even when the fall rate was as small as 4 Hz/csec; whereas, when the peak was at about two-thirds into the second syllable, a rate of 8 Hz/csec or greater was necessary for the same judgment.

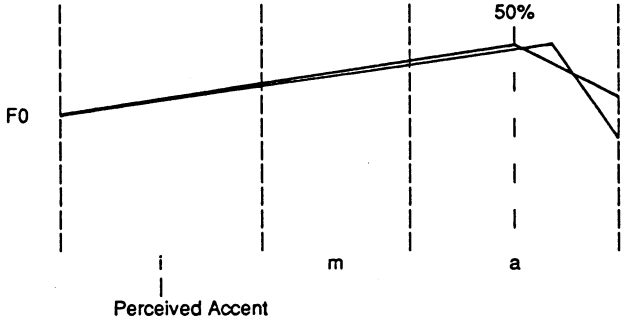


**Figure 3: F0 peak location and F0 fall rate**

The present study investigates whether or not the effect of F0 fall rate is observed in accent perception of English. The focus of English utterance, if there is any, is often expressed by placing the so-called contrastive accent on a certain syllable within the focused constituent (cf. Bolinger 1954, 1961, Halliday 1967, Chafe 1976, Lambrecht 1986). For example, the contrastive accent would be on *my* in "This is my net" if one were to answer the question "Is this her net?" On the other hand, in neutral (unmarked) intonation, e.g. responding to "What's this?", the nuclear accent occurs on the last syllable, *net*. Due to the coupling with final lowering, the fall rate must be greater if the nuclear accent is on the final syllable than elsewhere in the utterance (Mattingly 1968, Olive 1974, Maeda 1976)[3]. Thus, utterances with this condition are likely candidates in English for observation of the effect of F0 fall rate, if it should occur.

## 2. Perceptual Experiment

### 2.1 Method

Using a MITalk-based system, we synthesizedten variations of the English utterance, "This is my net", with different F0 fall rates on *net*. The duration and amplitude of each syllable were kept constant across the utterance-stimuli, and the F0 peak always occurred at the onset of /ɛ/ in *net*. The F0 contour of the utterances started at 121 Hz, linearly rose to 150 Hz at the onset of /ɛ/ in *net*, and ended at 102 Hz. The difference between the peak and the end was 48 Hz, and the duration of the fall was varied from 2 csec (= 20 msec) to 11 csec by a 1-csec step. The fall rate of each stimulus is shown in the following table.

| Stimulus | Rate (Hz/csec) | Stimulus | Rate (Hz/csec) |
|---|---|---|---|
| 1 | 4.4 | 6 | 8 |
| 2 | 4.8 | 7 | 9.6 |
| 3 | 5.3 | 8 | 12 |
| 4 | 6 | 9 | 16 |
| 5 | 6.9 | 10 | 24 |

**Table 1: F0 fall rates of stimuli**

Figure 4 illustrates the F0 contour, the duration, and the amplitude envelope of each segment of sample stimuli (1 and 10).

Thirty-four native speakers of American English participated in the experiment. First, they listened to the experiment instructions in synthetic speech in order to familiarize themselves with the synthetic voice. Then, the subjects were asked to judge whether each utterance was more appropriate to responding to "What's this?" (inducing the accent on *net* in "This is my net") or to "Is this her net?" (inducing the accent on *my*). Hereafter, the former will be referred to as *net-response*, and the latter as *my-response*. Each subject listened to two sets of the 10 stimuli which were randomized in different orders.

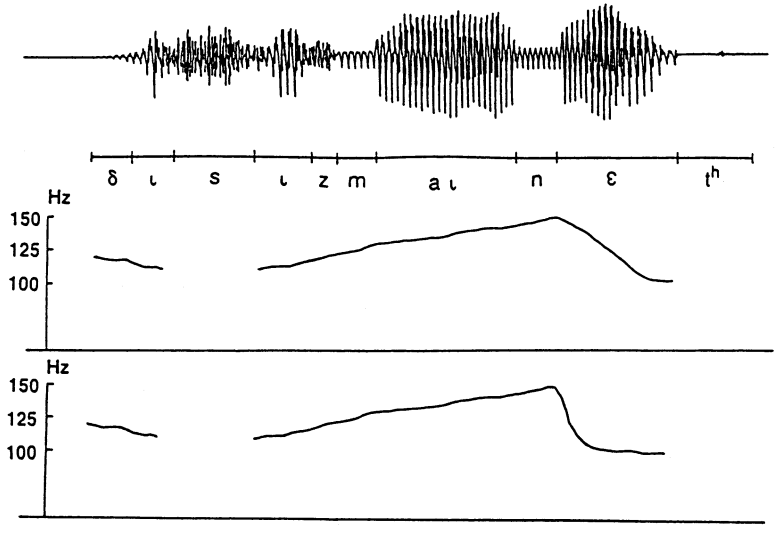


**Figure 4: Stimulus with 4.4 Hz/csec fall rate (top) and stimulus with 24 Hz/csec fall rate (bottom)**

## 2.2 Results and Discussion

Because subjects' responses were occasionally arbitrary, we counted only consistent judgments (i.e. the same judgment on both sets); the results are summarized in the following table. The first column indicates the fall rate, and the second indicates the percentage of consistent judgments (the number of subjects appears in parentheses); the third and fourth columns indicate the percentage of the *net-* and *my-responses*, respectively, against the total number of consistent judgments.[4].

| Rate (Hz/csec) | Consistent judgements | net-responses | my-responses |
|---|---|---|---|
| 4.4 | 91% (31) | 97% (30) | 3% (1) |
| 4.8 | 79% (27) | 96% (26) | 4% (1) |
| 5.3 | 79% (27) | 96% (26) | 4% (1) |
| 6 | 74% (25) | 84% (21) | 16% (4) |
| 6.9 | 65% (22) | 73% (16) | 27% (6) |
| 8 | 59% (19) | 79% (15) | 21% (4) |
| 9.6 | 62% (21) | 52% (11) | 48% (10) |
| 12 | 71% (24) | 50% (12) | 50% (12) |
| 16 | 71% (24) | 33% (8) | 67% (16) |
| 24 | 79% (27) | 41% (11) | 59% (16) |

**Table 2: Comparison of the net- and my-responses**

As shown in the second column, the consistency of subjects' judgments is highest at the two extreme fall rates: 91% when the rate is smallest (4.4 Hz/csec), and 79% when it is greatest (24 Hz/csec). The farther away the fall rate is from these two extremes, the fewer the consistent judgments. This fact indicates that if the F0 fall rate is significantly small or great, the listener can determine thelocation of accent consistently from F0 information alone, but if the rate is close to 8 Hz/csec, the F0 by itself is ambiguous as a cue to the accent location.

The third and fourth columns of the table show that there is asymmetry between the two judgments with greater and smaller fall rates. At 4.4 Hz/csec, 97% of the consistent judgments are net-responses. In contrast, at 24 Hz/csec, the subjects' judgments split between net-and my-responses. This implies that if the rate is small, the accent is perceived on the syllable where the actual F0 peak occurs, but if the rate is great, the accent is likely, but not necessarily, to be perceived on the preceding syllable.

Figure 5 plots the percentage of the responses with respect to the total number of consistent judgments. The general tendency for the greater fall rate to shift the perceived accent observed in this study is in accordance with the results of our previous experiments with Japanese accent. However, the proportion of the perceptual shift is smaller in English than in Japanese: the maximum percentage of the shift is close to 100% in Japanese, whereas it is only 67% (at 16 Hz/csec) in this experiment.[5].

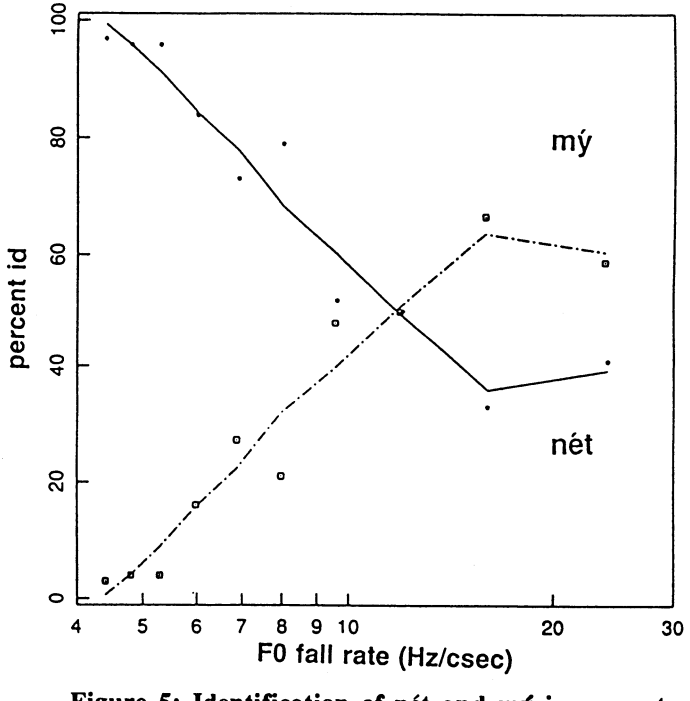


**Figure 5: Identification of nét and mý in percent**

In order to account for this difference, it is necessary to compare the prosodic structures of these two languages. Unlike Japanese, in which F0 is by far the most prominent indicator of accent (Weitzman 1969, Bechman 1986), English makes use of four orthogonal acoustic cues: F0, duration, amplitude, and spectral patterns. Generally, the most prominent indicator is F0, then duration, and then amplitude, as is reported by Fry. However, the ranking varies significantly among native listeners (Beckman 1986) as well as among the syntactic structures where the word in question appears (Nakatani and Aston 1978). It seems that accent is an impressionistic sum of cues, as Onishi claims, but the method of the summation is not uniform. Some native speakers of English have difficulty in detecting Japanese accent, which generally does not accompany longer duration nor higher amplitude.

In the present experiment, two subjects (of 34) never exhibited a shift of the perceived accent. It may be that F0 is a lower-ranking cue for these subjects, and therefore, perhaps it might have been more appropriate to exclude them from the data. However, because we have no evidence for their internalized ranking, we included these subjects in the analysis.

It is also necessary to point out the fact that the net-response is unmarked, whereas the my-response is highly marked; i.e., *net* is the neutral location to place the nuclear accent in the utterances without a special focus. Many subjects commented after the experiment that some stimuli were not as natural as they should be for "This is my nét" but that they nevertheless gave a net-response because the accent they perceived on *my* was not sufficient to carry a contrastive accent. Furthermore, in order to avoid the use of undefined notion *accent*, we selected pragmatic differences to distinguish two accentual patterns. This methodology placed an extra burden on some subjects, who claimed that the task was extremely difficult. They were not accustomed to thinking of the question to which a given utterance is appropriate. Therefore, if the subjects' task had been to judge the naturalness of utterances with the nuclear accent on *net*, the percentage of the net-response would have been much lower than the present results.

Taking into consideration these factors, we conclude that the shift of the perceived accent is observable in English. Similar to native Japanese listeners, native English listeners utilize the F0 peak location and the F0 fall rate as cues to determine the location of accent.

## 3. Implications

Understanding how the listener determines the place of accent is crucial in two areas: sound change and speech synthesis.

### 3.1 Sound change

Ohala (1981, 1983) claims that those sound changes attested in similar form in diverse languages result from errors of transmission of pronunciation from one speaker to another. The preconditions of such sound changes are some universal physical or physiological constraints which occur in present-day speech and are therefore available for investigation. For example, the change of labialized velars to labials is very common, but not vice versa. It has also been reported that the identification errors of /k/ as /p/ is much greater than /p/ as /k/ (Winitz et al. 1972).The representation of these sounds as velars and labials obscures the potential causes of the asymmetry in historical sound change and in the confusion matrices. We must know what we represent by these terms.

Like segmentals, suprasegmentals do change in the course of time. A language may lose an original quantity opposition, e.g. vowel length, or it may develop a new one. A language may lose or acquire distinctive tone: a formerly free and nondistinctive tone may become fixed and acquire the value of a boundary signal (Jeffers and Lehiste 1982).

In the case of Japanese accent, McCawley (1968) reports that the accent system of two syllable words in the Tokyo type dialect is likely to develop from the proto form by shifting accent one syllable to the right. Understanding how a high tone is perceived makes this claim very plausible. If the F0 peak occurred on the post-accent syllable, and the fall rate were not great enough, the accent would be perceived on that syllable. In Ohala's theory, this is a seed of sound change. Whether or not this change would spread must be accounted for in the sociolinguistic domain,

### 3.2 Speech synthesis

In synthesizing English utterances, the placement of the F0 peak of the nuclear accent continues to be controversial. Ashby (1978) claims that the nuclear accent occurs at a fixed location in the vowel regardless of the vowel length, whereas Steele (1986) claims that the peak location should vary according to the vowel duration. Furthermore, Pierrehumbert (1981) and Silverman (1987) report that the F0 peak location varies between the nuclear accent and other prenuclear accents. Their F0 algorithms for synthesizing English intonation place the peak earlier in the nuclear accent than in prenuclear accents.

Although Olive (1974) mentions that the fall rate of the nuclear accent must be greater than that of prenuclear accents, no study, to our knowledge, has considered the relationship between the peak location and the fall rate. Because the effect of F0 fall rate on accent perception is found in English, elaborated speech synthesizers of English utterances should take this relationship into consideration.

Let us reexamine the results in Table 2. Given the peak at the vowel onset in *net* and the 6.9 Hz/csec fall rate, 73% of the consistent judgments were net-responses. However, if we consider the total number of judgments, rather than the consistent judgments, less than half of the subjects (16 out of 34) perceived the accent on *net*. This indicates that, as the fall rate increases, approximately 7 Hz/csec is where F0 by itself ceases being a reliable cue to accent location. Moreover, if the rate is greater than 12 Hz/csec, other acoustic cues (i.e. longer duration, higher amplitude, and different spectral patterns) may compete against the F0 cue for those who rely heavily on F0 in determining accent. We, therefore, suggest that in order to avoid the effect of F0 fall on English accent perception, the fall rate should not exceed 12 Hz/csec when the peak location is at the onset of the vowel of the syllable which carries the nuclear accent.

## 4. Conclusion

In the present study, we found that English also manifests the effect of F0 fall rate on accent perception observed in Japanese. The results show that manipulating the fall rate alone can cause a perceived accent to be shifted in English utterances. Because English provides other acoustic cues in addition to F0, the occurrence of the perceptual accent shift is less frequent than in Japanese.

Accent and intonation, however they are defined, are essential parts of language. It is extremely difficult for non-native speakers to acquire normal

accentual patterns. One reason is that the composition of accent is complex and language-specific even though the components are chosen from the pool of features which are available universally. Therefore, those components which play a significant role in accent placement must be stated explicitly in the description of languages.

## Notes

[1]An earlier version of this paper was presented at the 120th Meeting of the Acoustical Society of America, November 26-30, 1990, San Diego, California. We would like to thank the following individuals for comments on various stages of this work: Michelle Caisse, John Cherry, Carlos Gussenhoven, Michael O'Mailly, John Ohala, Raymond Weitzman, and Helen Wheeler.

[2]Fujisaki et al. (1976) have suggested that the desynchronization of the F0 peak and the syllable boundary in acoustic data is not psychologically real but a mere reflection of different processing time between detecting F0 changes and segmental boundaries. Detecting F0 changes is faster than detecting segmental boundaries, and thus, they synchronize in perception. Javkin (1976), and Maddieson (1976) conducted experiments to determine when F0 changes and segmental boundaries are recognized, but their results do not provide conclusive evidence for this hypothesis.

[3]The fall rates which we calculated from the F0 data of English utterances in Maeda (1976) show that they are greater in sentence-final position (10.2-14.5 Hz/csec) than in other positions (8.4-13.3 Hz/csec).

[4]The result of a chi-square test shows that the differences between my- and net-responses are significant at the 1% level.

[5]Michael O'Mailly pointed out that the relatively high percentage of net-response at 24 Hz/csec might be due to the defect of the stimulus, i.e. the stimulus might not have the fall rate of 24 Hz/csec (emwhich frequently is the case with synthesizers based on certain algorithms. We rechecked the the F0 fall rates of the all stimuli, and confirmed the accuracy of their fall rates.

## References

Ashby, M. 1978. A study of two English nuclear tones.*Language and Speech* 21, 326-336.

Beckman, M.E. 1986. *Stress and Non-stress Accent* .Dordrecht, Holland: Foris Publication.

Bolinger, D. 1954. English prosodic stress and Spanish sentence order. *Hispanica* 37, 152-56.

Bolinger, D. 1961. Contrastive accent and contrastive stress. *Language* 37, 83-96.

Chafe, W. 1976. Givenness, contrastiveness, definiteness, subjects, topics and point of view. In C. Li (ed.) *Subject and topic*.New York: Academic Press.

Fry, D.B. 1958. Experiments in the perception of stress. *Language and Speech* 1, 126-152.

Fujisaki, H., H. Morikawa, and M. Sugito. 1976. Temporal organization of articulatory and phonatory controls in realization of Japanese word accent. *Annual Bulletin, RILP, University of Tokyo* 10, 177-90.

Halliday, M.A.K. 1967. Notes on transitivity and theme in English, part II. *Journal of Linguistics* 3, 199-244.

Hasegawa, Y. and K. Hata. 1988. Delayed pitch fall in Japanese.*JASA* Suppl.1.83, S29.
Hata, K. and Y. Hasegawa. 1988. Delayed pitch fall in Japanese: a perceptual experiment. *JASA* Suppl.1.84, S156.
Javkin, H.R. 1976. Auditory basis of progressive tone spreading. *JASA* Suppl.1.60, S45.
Jeffers, R.J. and I. Lehiste. 1982. *Principles and methods for historical linguistics.* Cambridge: MIT Press.
Lambrecht, K. 1986. Topic, focus, and the grammar of spoken French. Ph.D. dissertation. University of California, Berkeley.
Lehiste, I. and G.E. Peterson. 1961. Some basic considerations in the analysis of intonation. *JASA* 33, 419-25.
McCawley, J. 1968. *The phonological component of a grammar of Japanese.* The Hague: Mouton.
Maddieson, I. 1976. Tone spreading and perception. *JASA* Suppl.1.60, S45.
Maeda, S. 1976. A characterization of American English intonation. Ph.D. dissertation, MIT.
Mattingly, I. 1966. Synthesis by rule of prosodic features. *Language and Speech* 9, 1-13.
Nakatani, L.H. and C.H. Aston. 1978 Acoustic and linguistic factors in stress perception. Unpublished manuscript, Bell Laboratories.
Neustupný, J.V. 1966. Is the Japanese accent a pitch accent? *Onsei-Gakkai Kaihoo* 121. Reprinted in M. Tokugawa (ed.), *Akusento.* Tokyo: Yuuseidoo. 1980. 230-239.
Ohala, J.J. 1981. The listener as a source of sound change. *CLS* 17, 178-203.
Ohala, J.J. 1983. The direction of sound change. In A. Cohen and M.P.R.v.d. Broecke (eds.) *Abstracts of the 10th International Congress of Phonetic Science.* Foris: Dordrecht. 253-58.
Olive, J. 1974. Speech synthesis by rule.Speech Communication Seminar, Stockholm, Aug. 1-3, 1974.
Onishi, M. 1942. Kokugo akusento-ron. In M. Togo (ed.) *Nihongo no akusento.* Tokyo: Chuo-koron. 15-26.
Pierrehumbert, J. 1981. Synthesizing intonation. *JASA* 70, 985-95.
Silverman, K.E.A. 1987. The structure and processing of fundamental frequency contours. Ph.D. dissertation, University of Cambridge.
Steel, S.A. 1986. Nuclear accent F0 peak location: effects of rate, vowel and number of following syllables. *JASA* Suppl.1.80, S51.
Sugito, M. 1972. Ososagari-koo: dootai-sokutei ni yoru nihongo akusento no kenkyuu (Delayed pitch fall: an acoustic study). *Shoin Joshi Daigaku Ronshuu* 10. Reprinted in M. Tokugawa (ed.), *Akusento* (Accent). Tokyo: Yuuseidoo. 1980. 201-229.
Weitzman, R. 1969. Japanese accent: an analysis based on acoustic-phonetic data. Ph.D. dissertation, University of Southern California.
Winitz, H., M.E. Scheib, and J.A. Reeds. 1972. Identification of stops and vowels for the burst portion of /p,t,k/ isolated from conversational speech. *JASA* 51, 1309-17.