Title
Truth in Concentration in the Land of (80/20) Laws

Journal
Marketing Science, 12(2)

ISSN
0732-2399

Authors
Schmittlein, David C
Cooper, Lee G
Morrison, Donald G

Publication Date
1993-05-01

DOI
10.1287/mksc.12.2.167

MARKETING SCIENCE
Vol. 12, No. 2, Spring 1993
Printed in U.S.A.

TRUTH IN CONCENTRATION IN THE LAND OF (80/20) LAWS

DAVID C. SCHMITTLEIN, LEE G. COOPER
AND DONALD G. MORRISON
University of Pennsylvania
University of California, Los Angeles
University of California, Los Angeles

Among the more prominent truisms in marketing are 80/20 type laws, e.g., 20 percent of the customers account for 80 percent of the purchases. These kinds of statistics indicate a certain degree of *concentration* in customer purchases; i.e., the extent to which a large portion of the product's total purchases are made by a small fraction of all customers. Such concentration levels, suggesting that markets can be segmented in various ways, are often reported in basic marketing texts.

We show that a *meaningful interpretation* of these concentration statistics is not nearly as easy or immediate as it is to compute them. The key factors influencing the degree of *apparent concentration* in purchases are reviewed, and we present a modeling approach for estimating the true level of *relevant concentration* among customers.

(Buyer Behavior; Estimation and Other Statistical Techniques; Market Structure; Measurement)

Introduction

In Twedt (1964) the question "How important to marketing strategy is the 'heavy user'?", was asked. Twedt addressed this question by looking at *observed* concentration statistics obtained from 12 months of *Chicago Tribune* panel data collected in 1962. He presented data on 18 product categories all in the format of the following two examples.

		Users	
	Nonusers	Light Half	Heavy Half
Ready-to-eat Cereals	4%	13%	87%
Canned Hash	68%	14%	86%

Thus in ready-to-eat cereals only 4 percent of the households made no purchases. Of the 96 percent of households that purchased cereal, those purchasing more than the median amount (i.e., the "heavy half") accounted for 87 percent of all purchases. Of course the "light half" made up the remaining 13 percent of purchases. The canned hash numbers are interpreted analogously.

0732-2399/93/1202/0167$01.25

Note that *among users*, *observed* concentration is virtually identical for cereal and canned hash. However, canned hash has 17 times more nonusers! On average, cereal is also purchased much more frequently than canned hash. What is a meaningful and useful interpretation of the kinds of concentration statistics above? We believe that these statistics, standing alone, provide a very incomplete (and sometimes misleading) picture of the concentration in purchasing. The objective of this paper is to show the reader how to develop more relevant concentration figures from the panel data used by Twedt in the early 1960s. (Of course, this type of data is much more easily available today.)

By "relevant" we mean concentrations that are based solely on the across-household variability in the underlying (unobservable) purchasing rates. This variability is the only source of differences across households that will actually hold up over time. Twedt's statistics are, of course, contaminated by the within-household variance in the number of units actually purchased in some particular time period.

By way of example, consider the soup purchases of two separate households denoted A and B. Household A's long-run purchase rate is three units/month, and the rate for B is six units/month. Of course, in most months the observed number of units actually bought will deviate from this long-run average, due to promotions, variety seeking, weather/seasonality, and so on. By analogy with some relevant psychometric literature (Lord and Novick 1968), the long-run purchase rate can be thought of as a "true" score or latent trait for each household. The "observed" score in any one-unit time period is then this latent trait plus a random "error" component.

We will occasionally refer to the degree of concentration in these true scores (long-run purchase rates) as "true concentration." Besides the obvious connection to "true" scores, we have two reasons for this appellation. First, as noted earlier, this is the only notion of concentration that will hold true over the longer term, e.g., for targeting households that will persistently be heavy users. Second, by estimating the concentration in long-run purchase rates, we can predict quite accurately the "observed" level of concentration (i.e., concentration in the number of observed purchases) in a time period of *any* given length. We do not mean to suggest that these "observed concentration" levels are "un-true," but rather that they can be better interpreted.

Table 1 shows clearly some of the problems in directly interpreting observed concentration levels. The top half shows the percent of units accounted for by the top 20 percent of households, for four product categories, as a function of the amount of time these households were observed. We see a substantial and systematic decline in observed concentration as the time period increases. Indeed, for yogurt purchases that decline has not ceased even in two years of purchasing. How is one to list a "true" level of concentration for these product categories?

The bottom half of Table 1 lists the observed concentration *among users* (as in Twedt's statistics), i.e., among those households that happen to have bought at least once in the time period analyzed. Here the pattern is sometimes an *increase* in observed concentration with time (catsup, yogurt), sometimes a slight decrease in concentration with time (soup), and sometimes no change at all with time (detergent). For catsup, yogurt, and soup purchasing among users, we again have the question, "What is the true level of concentration among users?" Indeed, here we have another unanswered question, namely "Which product categories are more concentrated, and which less?" During a typical six-month period, soup appears more concentrated (among users) than catsup. But during a two year period, catsup's observed concentration is greater than that for soup.

We will see that "truth in concentration" depends in some rather subtle ways on:

- The model assumed—in particular, whether it allows for *true* nonusers—as opposed to individuals who happen not to have bought during the time period in which observations were taken.
- What we do with the *observed* nonusers.

TABLE 1

The Percent of Total Purchases Due to the Top 20% of Customers

	Top 20% of the Entire Population			
Time Period of Observation	Catsup	Detergent	Yogurt	Soup
1 month	86	65	97	56
3 months	63	55	85	49
6 months	56	52	80	49
9 months	54	50	78	48
1 year	53	50	78	46
2 years	52	49	73	45
	Top 20% of "Users"			
Time Period of Observation	Catsup	Detergent	Yogurt	Soup
1 month	32	44	49	48
3 months	42	47	58	47
6 months	46	48	62	48
9 months	48	47	63	48
1 year	49	47	65	46
2 years	50	48	65	45
Average Number of Purchases per Household per Year	4	11	16	49

N = 3836 households.

- The length of the observation period.
- The average purchasing rate, across households.
- The heterogeneity of purchasing rates across households.
- The degree of regularity of the household interpurchase times, e.g., "random" (exponentially distributed) or more regular (Erlang 2 distributed) times.
- The effect on concentration of using numbers of purchase occasions versus total amount purchased versus dollar amount spent. (This last point will not be addressed until we present the empirical findings.)

Collecting Twedt-type concentration statistics is useful and will give managers some insights. On the other hand, naive comparison of these statistics for (say) frozen orange concentrate and apple juice can result in very misleading "apples and oranges" comparisons. There are two key problems with this idea of simply "looking at" the observed concentration levels. First, as we saw above, the apparent concentration varies depending on the amount of time for which customers are observed. Second, the well-known "regression to the mean" effect (cf. Morrison and Schmittlein 1981) implies that any particular group of customers in which purchases were observed to be concentrated (e.g., the "heavy half") will fail to provide the same concentration in purchases in any subsequent time period.

In contrast, assessing concentrations as proposed in this paper

- is the only general approach that is time-invariant, i.e., the expected measured concentration does not depend on how long we watch;
- provides a measure of the concentration that will be observed "in the long run"; and,
- allows the manager to predict the degree of concentration that will be provided in each future period by any set of customers (e.g., this period's heavy half).

In this way our approach will yield an "apples to apples" comparison. Perhaps more importantly, we will highlight some of the subtleties that are almost never discussed when

looking at heavy half statistics or quoting that old saw, the 80/20 Law (20 percent of the customers account for 80 percent of the sales). That is, concentration statistics are influenced by various factors, and proper interpretation requires an understanding of these factors. Our intended contributions are to enhance understanding of Twedt-type concentration statistics and provide a methodology for computing more illuminating concentration statistics. In this paper we will not tell managers *how* to target the important heavy users. But we will show the manager *where* (i.e., in which product categories or brands) he or she is most likely to encounter true heavy users who are most worthwhile to target.

Finally, this paper has an empirical flavor. Data on tuna, catsup, detergent, orange juice, soup, toilet tissue and yogurt are presented. These analyses are done at both the category and brand level. Some rather surprising results occur, which cause us to rephrase Twedt's 1964 question: "How many and just how heavy are the important users?"

True (Long Run) Concentration in the NBD World

We will illustrate our basic idea with the well-known NBD model of product purchase frequency. This parsimonious model can be used to predict a variety of market statistics such as the distribution of purchase frequencies across households, the average number of purchases per buyer, and the market-penetration level. It also predicts how these quantities will vary depending on the duration of the time period being considered. The NBD model has extensive empirical support (Ehrenberg 1988; Morrison and Schmittlein 1981), although there are some specific features of the purchase process that it does not capture completely (Morrison and Schmittlein 1988). (We examine a generalization of the NBD to handle these features in the main empirical results to follow.)

The NBD incorporates the two main sources of variation in purchase frequencies discussed in the Introduction:

(1) The variation, within household over time, in the actual frequency of purchase (i.e., around the household's long-run purchase rate). This is the source of concentration in Twedt's statistics that will not lead to any sustained differences across households (i.e., enabling behavior-based segmentation of customers), and makes the simple observed statistics cited by Twedt and others dependent on the length of time period examined (since the size of this within-household variance depends on the time-period duration).

(2) The variation across households in their long-run purchase rates. This is the source of what we term "true concentration," for the reasons outlined earlier.

For these two sources of variation, the NBD model makes the following two assumptions, respectively:

(1) Each household has an underlying purchasing rate λ, where conditional on λ the number of units actually purchased X, is Poisson distributed with mean λ.

(2) Household purchasing rates are distributed gamma (r, α) across the population of households.

Mathematically we have:

for household i with purchasing rate λ_i,

$$P_p(X_i = x|\lambda_i) = \frac{(\lambda_i t)^x}{x!} e^{-\lambda_i t}, \qquad x = 0, 1, 2, \ldots .$$

(This assumes the time period of observation is of length t.) Across households, the density function for purchasing rates is

$$g(\lambda|r, \alpha) = \frac{\alpha^r}{\Gamma(r)} \lambda^{r-1} e^{-\alpha\lambda}, \qquad \lambda > 0; \quad r, \alpha > 0. \tag{1}$$

When the individual household purchases are aggregated, we obtain the negative binomial distribution:

$$P_{\mathrm{NBD}}(X = x) = \int_0^1 P_p(X = x|\lambda)g(\lambda|r;\alpha)d\lambda$$

$$= \frac{\Gamma(r+x)}{x!\Gamma(r)}\left(\frac{\alpha}{\alpha+t}\right)^r\left(\frac{t}{\alpha+t}\right)^x, \qquad x = 0, 1, 2, \ldots. \tag{2}$$

The Twedt-type observed concentration statistics in an NBD World are based on the X's generated by (2). Of course the "true concentration" is a function of the distribution of the households' long-run purchasing rates λ.

The Lorenz Curve and Customer Concentration

Although most readers are acquainted with the Lorenz curve, a brief review may be helpful—particularly since the Lorenz curve is by far the most crucial construct in this paper. (For a more detailed coverage of this and other concentration indices, see Cowell 1977, Curry and George 1983 or Schmittlein 1988.) If we sort the customers from those making the fewest purchases to those making the most, and plot the cumulative number of purchases, the Lorenz curve $L(p)$ is the proportion total volume (total purchases, in our case) accounted for by those households in the p^{th} percentile or less. Figure 1 contains everything that we need. The curve OCB is the Lorenz curve. The particular point C means that (in this example) the 80 percent of the households that purchase at the 80th percentile or less account for only 20 percent of the total purchases.

If the Lorenz curve had been the 45 degree straight line OB, then every household would have purchased *exactly* the same amount, i.e., the bottom half of the households would account for half of the total purchases. If, on the other hand, the Lorenz curve had been the right angle OAB, then all households except one would have made no purchase and this remaining household would have purchased everything. An OB Lorenz curve implies no concentration of purchasing (everyone purchases the same amount),

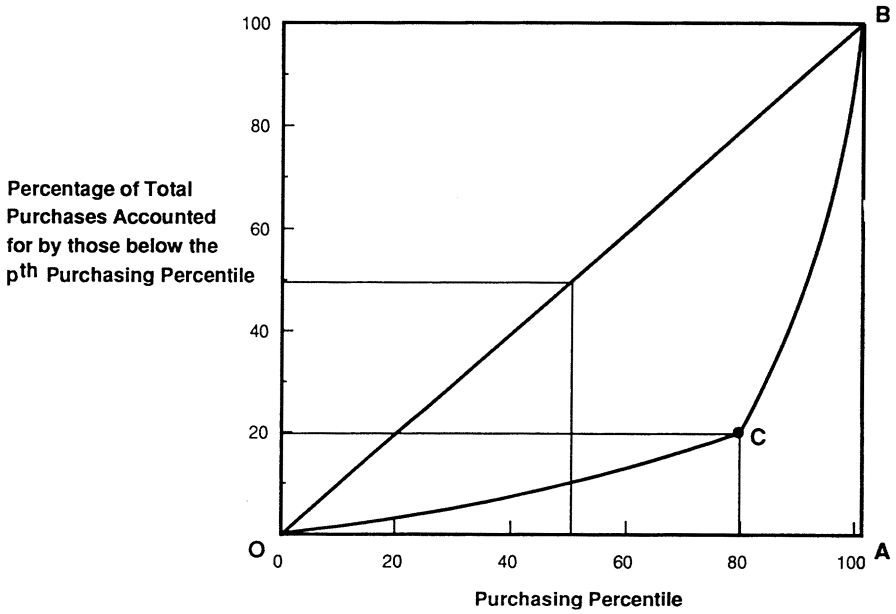


FIGURE 1. The Lorenz Curve

while an *OAB* Lorenz curve is the ultimate in concentration. In general, the more "bowed" the actual Lorenz curve *OCB* is, the higher the degree of concentration.

Light and Heavy Halves as One Point on the Lorenz Curve

In our purchasing context, the light half of households accounts for

$$L_x(0.5) = \frac{\sum_{x=0}^{\text{Median}} xP(X = x)}{\sum_{x=0}^{\infty} xP(X = X)} = \frac{1}{\text{Mean}} \sum_{x=0}^{\text{Median}} xP(X = x). \tag{3}$$

The subscript "x" indicates that the Lorenz curve is based on the observable number of purchases x. The heavy half then accounts for

$$(1 - L_x(0.5)) \times 100 \text{ percent}$$

of the total purchases.

Observed Light Half in an NBD World

The observed light half proportion is just a particular point on the Lorenz curve associated with the observed NBD mixture distribution (2). Although the light half and heavy half will still play a role in this paper, our main focus will be on the complete Lorenz curve, $L(p)$, for $0 < p < 1$. The 80/20 "law" referred to in the introduction says that "the top 20 percent of the customers account for 80 percent of the sales" (which also means that the bottom 80 percent account for only 20 percent of the sales). The 80/20 Law thus says $L_x(0.8) = 0.2$.

True Lorenz Curve in an NBD World

Since each household has an unobservable true purchasing rate λ, the true Lorenz curve will be analogous to (3)—but based on the unobservable mixing distribution $g(\lambda \mid r, \alpha)$. Letting $G(\lambda \mid r, \alpha)$ be the CDF corresponding to the density function g, we have

$$L_\lambda(p) = \frac{1}{E[\lambda]} \int_0^{G^{-1}(p)} \lambda g(\lambda) d\lambda. \tag{4}$$

Again, the "λ" subscript indicates that the Lorenz curve is based on the unobservable mixing distribution on λ. Also, (4) holds for any mixing distribution, but in this paper we will focus on the gamma distributions commonly used in modeling purchases.

Estimating $L_\lambda(p)$

In an NBD world, the true Lorenz curve (i.e., the curve related to long-run purchase rates) $L_\lambda(p \mid r, \alpha)$ is a function of the percentile of interest p, and the two parameters of the gamma mixing distribution, r and α. The NBD distribution will be fit to the observed histogram of the number of purchases across households. This fitting will involve estimating r and α (via maximum-likelihood in our empirical results to follow). With these estimates, $\hat{r}$ and $\hat{\alpha}$, we can then sweep out the complete Lorenz curve $L_\lambda(p)$, letting p range from 0 to 1. The corresponding Lorenz curve for the observed purchases $L_x(p)$ can be calculated from the actual histogram of purchases (which we will do) or from the "smoothed" NBD predicted value for the histogram.

The true Lorenz curve $L_\lambda(p)$ can be derived by substituting Equation (1) into (4). Letting $F(\lambda \mid r, \alpha)$ denote the C.D.F. for the gamma distribution (1) and F^{-1} be the inverse function of F, Equation (4) becomes

$$L_\lambda(p) = \frac{\alpha}{r} \int_0^{F^{-1}(p \mid r, \alpha)} \frac{\alpha^r \lambda^r}{\Gamma(r)} e^{-\alpha\lambda} d\lambda = \int_0^{\alpha F^{-1}(p \mid r, \alpha)} \frac{\lambda^r}{\Gamma(r+1)} e^{-\lambda} d\lambda. \tag{5}$$

Using (1) it is easily shown that

$$F^{-1}(p|r, \alpha) = \alpha^{-1}F^{-1}(p|r, 1) \tag{6}$$

and substituting (6) in (5), the desired true Lorenz curve for purchase rates is

$$L_\lambda(p) = F(F^{-1}(p|r, 1)|r + 1, 1). \tag{7}$$

where $F(\ \ |r, \alpha)$ is, again, the gamma distribution C.D.F.

One point about (7) is worth noting: the true (long run) concentration level $L_\lambda(p)$ depends *only* on the shape parameter r of the gamma mixing distribution, and not on the scale parameter α. Thus r itself can be viewed as an overall inverse measure of the concentration in purchase rates across households, since $L_\lambda(p)$ in (7) increases as r increases. (For another way to see this role of r, note that $1/r$ is the squared coefficient of variation in purchase rates across households; see Appendix.) We will have more to say about the implication of high versus low r-values in the empirical results below.

For purchases made in accordance with the NBD model, the observed Lorenz curve $L_x(p)$ is always below the Lorenz curve based on long-run purchase rates $L_\lambda(p)$. This discrepancy decreases, of course, as the length of the observation period increases (i.e., as observed purchases for each household settle down to their long-run rate). That is, the observed $L_x(p)$ increases toward $L_\lambda(p)$. This *decrease* in "observed concentration" toward "true concentration" (note that increasing $L_x(p)$ for each given p means decreased concentration) is what we saw previously in the top half of Table 1.

This time effect confounds any assessment of concentration based on the observed statistics alone. A model for purchase patterns—such as the NBD discussed earlier—is required in order to assess both the "true" concentration and the concentration that *will* be observed in any time period of arbitrary duration. We turn next to the selection of the "most appropriate" model for this task.

Modifying the NBD and "What to Do with the Zeroes?"

The NBD model has two somewhat unappealing properties (random purchasing and no nonusers over the long run)—the latter is particularly troublesome within this concentration in purchasing arena. There is also the nagging issue of what, if anything, should we do with the consumers who actually make zero purchases. Twedt ignores them when calculating light-half and heavy-half statistics. But surely some of these "zeroes" are *not* hard-core nonpurchasers who will never buy in the product category. Intuitively therefore some, but perhaps not all, of the "zeros" should be "counted" in the concentration statistics.

Nonuser NBD Model

In the Introduction, we saw that 68 percent of the households purchased no canned hash during the twelve months of observation. Surely some—perhaps most—of these households will *never* buy canned hash. Of course, some of these "zeros" will eventually purchase, although their underlying purchasing rates λ are obviously quite small. The Nonuser NBD model (NUNBD), first introduced by Morrison (1969), is the natural extension of the NBD that incorporates these hard-core nonusers. Quite simply, we let a proportion q of the households have a purchasing rate $\lambda = 0$. The remaining $1 - q$ proportion of households have their λ values distributed gamma (r, α). In other words, we are merely putting a mass point of size q at the $\lambda = 0$ point of the unobservable mixing distribution.

The true concentration statistics for this NUNBD model can be calculated in two different ways. First, we include the mass point of size q at $\lambda = 0$ in the integral (4).

Thus, if $q = \frac{1}{2}$, the true heavy half of the total population would account for 100 percent of the total purchases—regardless of the gamma parameters r and α.

The second, and our preferred, method is to say:

(1) A proportion q of the households do not purchase this product category, and

(2) of the remaining $1 - q$ households, the Lorenz curve is calculated by (4) *excluding* the mass at $\lambda = 0$.

This method thus assesses the concentration among *true users* of the category, not concentration in the total population of users-plus-nonusers. The necessary mathematics for the NUNBD are given in the Appendix.

The Observed Lorenz Curve in an NUNBD World

If the "zeros" are not counted as in Twedt's statistics, then ready-to-eat-cereal and canned hash show very similar concentrations. If all of the "zeros" are included, then these two product categories look very different with respect to concentration. In an NUNBD world, both of these extreme counting rules fail to reflect households' underlying purchase propensities, since some of the "zeros" are hard-core nonusers with $\lambda = 0$ while others have $\lambda > 0$ and just happened not to purchase in this particular period. This latter phenomenon is particularly prevalent when the period of observation is short.

In contrast, anyone who explicitly uses the NUNBD model and estimates q, α, and r can report

- the "true" proportion q of hard-core nonusers, and
- a graph of the "true" Lorenz curve based on the gamma (r, α) mixing distribution for purchase rates λ among the remaining proportion $1 - q$ of users.

We will have more to say on these issues when the empirical results are presented.

More Regular than "Random" Purchasing

The NBD model assumes that households purchase in a Poisson manner. That is, the interpurchase times follow the memoryless exponential distribution—whose other properties include a constant hazard rate and a coefficient of variation of one. Numerous empirical studies (Herniter 1971, Lawrence 1980, Gupta 1988) show that observed interpurchasing times are more regular than implied by Poisson purchasing. Wheat and Morrison (1990) give a detailed analysis of how to interpret the various statistical methods for assessing regularity. We should also point out that Kahn et al (1986) and particularly Kahn and Morrison (1989) show that the household-level purchasing process may be considerably more "random" than implied by the Herniter, Lawrence, Gupta, and similar empirical studies. Nevertheless, the issue of household purchasing regularity is particularly crucial within our true-versus-observed Lorenz-curve setting.

Consider a population of households that purchase in a deterministic-clockwork manner. One of these households may buy coffee on the *average* of once per week and thus have a purchase event *every* seven days. Another household could have an average consumption of 26 units per year and would thus generate a purchase event every 14 days. A little reflection shows that except for end effects, i.e., the observation period is not an integer multiple of the purchase cycle, each household's observed number of purchases X equals its true average number of purchases λ. In the clockwork world, the observed Lorenz curve is the same as the true Lorenz curve.

Of course, such a degree of regularity simply does not exist. On the other hand, if purchasing is more regular than exponential, then the within-household variation in the number of units purchased is less than implied by the Poisson component of the NBD

model. Thus the true Lorenz curve will still be above the observed one—but it will not be as far above as implied by the NBD model.[1]

Thus far, we have attempted to cover the important issues of:

- true versus observed Lorenz Curves,
- model assumptions and their implications and
- the crucial—yet tricky—role played by the observed zeroes.

All of these points will now be illustrated in a set of rather extensive empirical findings. Before doing so, however, we should repeat our last "bulleted" point in the introduction.

Counting Events: Purchase Occasions versus Purchase Volume versus Total Dollars Spent

The modeling issues are enormously simplified by counting the number of purchasing *occasions* at the household level. The usual integer valued renewal processes can then be used to model household purchasing. But, of course, not all households always buy exactly *one* unit of the *same* package size at the *identical* price. Unfortunately, the latter two metrics—total pounds and total dollars—are more relevant to the managers using concentration statistics. Fortunately, the more parsimoniously modeled number of purchase occasions is typically very strongly related to the amount purchased, whether measured in pounds or dollars.

To illustrate this we report concentration statistics for both purchase frequencies and amount bought (in ounces) in Table 2. Households were tracked over a three-year period, and only those who bought the product during this period are included. Thus the table reports the concentration among apparent users discussed in the Introduction. The results show only a small discrepancy between concentration in number of purchases made, and total number of ounces bought. For example, the top 20 percent of yogurt buyers in terms of purchase frequency made 64 percent of the purchases. The top 20 percent in terms of amount bought account for 63 percent of the total ounces purchased. The concentration levels in the table never differ by more than five percentage points.

While the discrepancy in concentration levels is not large, the pattern is interesting. The concentration in purchase amounts was always slightly less than the corresponding concentration in purchase frequency. Note that the variation across households in their underlying consumption rates will determine (over the three year period analyzed) the variation in total amount bought (in ounces). If stores offered only one size package for the product, the concentration in purchase frequencies would of necessity equal that in amount bought (since the former would be a constant multiple of the latter).

But with different package sizes available, some households will elect to shop relatively often (purchasing—for them—relatively small sizes per trip) and others will shop more rarely (Kahn and Schmittlein 1989). This variation across households in shopping trip frequency will induce an *additional* variation in purchase frequencies, relative to that in purchase amounts. The result will be more concentration (across households) in frequencies than in amounts. One would expect such an effect to be most noticeable for products which are both easily inventoried in the household (stock-up goods) and available

[1] Chatfield and Goodhardt (1973) developed a model which captures the more-regular-than-Poisson purchasing. They called it the Condensed Negative Binomial Distribution (CNBD) model. The purchasing rates, λ, across households remain gamma distributed. However, the interpurchasing times now have an Erlang-2 distribution.

The true Lorenz curve estimated assuming a CNBD world will be *above* the observed Lorenz curve since there is still random variation in household purchases about the true household means. Similarly the CNBD true Lorenz curve will be *below* the corresponding NBD-model's estimate of the true curve, since the household-level random variation in the CNBD world is less than in the NBD world. To put it more succinctly, the CNBD scenario is somewhere between the random-purchasing NBD assumption and the ultimate in regularity, the deterministic-clockwork world.

TABLE 2

Concentration in Number of Purchases Versus Concentration in Amount Purchased (Three-Year Purchase History)

Cumulative % Users	10	20	30	40	50
Yogurt, Number of Purchases	45	64	77	85	91
Yogurt, Amount Purchased	44	63	75	84	90
Number of Users: 2599					
Catsup, Number of Purchases	34	54	67	77	86
Catsup, Amount Purchased	32	51	64	75	83
Number of Users: 3036					
Detergent, Number of Purchases	32	49	62	72	81
Detergent, Amount Purchased	27	45	59	70	79
Number of Users: 3091					

in a wide variety of sizes. Indeed, the largest such discrepancy in concentration levels was found for detergent purchases.

Overall, we conclude that the within-household and across-household variation in "amount" is reasonably well captured by the corresponding variability in "occasions." Thus the occasions-based NBD type models are appropriate for assessing "truth in *relevant* concentration."

Short-run and Long-run Concentration: Some Empirical Results

Most readers have been exposed to the standard Lorenz curve used in beginning economics textbooks; that is the form we used earlier. Namely, the horizontal axis defines the *bottom* x percent. In this empirical section, we ask the readers' indulgence as we "invert" the usual Lorenz curve. Marketing managers tend to say something like, "The *top* 20 percent of consumers account for . . .". Thus, we have chosen to present our results that way. Looking at Figure 1, point C on the inverted Lorenz curve will be reversed to a horizontal axis value of 20 and a vertical axis value of 80. Alternately, our inverted Lorenz curve is merely the reflection of the original Lorenz curve about the 45 degree line OB.

NUNBD Including Observed Zeroes

Since we wish to allow for the presence of some "hard-core nonusers," i.e., consumers with purchase rate $\lambda = 0$, then the NUNBD is an appropriate model. Table 3 gives the results for canned tuna fish. The two NBD parameters, $r = 1.383$ and $\alpha = 0.677$, were calculated from the distribution of purchase frequencies across households during a three-month period, as was the estimated proportion $q = 0.123$ of households having $\lambda = 0$. Thus we estimate that 12 percent of the population will never buy.

The actual concentrations in Table 3 include the zero purchases, e.g., the top 20 percent are the 20 percent rank ordered with respect to number of purchases of *all* households. The model predictions are for the expected observed concentrations, given the above NUNBD parameters.[2] The true long-run concentration is calculated from (4) based again on the above values for r and α.

[2] These NUNBD model expected observed concentration levels for various time periods are obtained by substituting Equation (2) in Equations (A1) and (A2), varying the length of time t in Equation (2). Doing so provides the probability distribution $P[X = x]$. This distribution, in turn, allows us to compute concentration levels via the discrete-distribution form of Equation (4) (with x substituted for λ) as illustrated in Equation (3). Since the NUNBD model predictions in Tables 3–8 are close to the observed concentration level in one-month and three-month time periods, we have some faith in the use of this model to calculate concentration curves for other time periods (including the long-run).

TABLE 3
Concentration among All Households, I.E., Including "Zeroes"
NUNBD *Model Versus Observed*
Tuna Purchases

Cumulative % Households	10	20	25	30	40	50
1-Month Actual % Purchases	44.73	69.02	77.26	85.51	100	100
1-Month Model % Purchases	47.13	71.05	79.42	87.79	100	100
3-Month Actual % Purchases	38.19	59.52	67.90	75.01	86.18	92.93
3-Month Model % Purchases	38.15	59.73	68.10	75.02	86.19	92.92
6-Month Model % Purchases	35.25	55.91	63.97	70.95	81.95	90.05
Estimated Long-Run % Purchases	32.02	51.44	59.09	65.92	77.15	85.59

NUNBD Parameters: $r = 1.383$; $\alpha = 0.677$; $q = 0.123$.

First note that as the period of observation *increases*, the observed concentration based on actual purchases *decreases*. This happens because, as time increases, individual households' actual purchases X become (in percentage terms) closer to their true rate λ.

Second, the actual concentrations are less than implied by the 80/20 Law. Even with the one-month time period, the top 20 percent account for only 69 percent of the purchases. If we *excluded* the zero purchases, this 69 percent figure would be even *less*. This occurs because the shape parameter $r = 1.383$ of the gamma mixing distribution $g(\lambda)$ is atypically large. The larger the value of r, the more homogeneous the population of consumers. That is, the coefficient of variation

$$\frac{\sqrt{\text{Var}\,[\lambda]}}{E[\lambda]} = \frac{1}{\sqrt{r}}$$

is smaller as r becomes larger. Many product classes have r values in the 0.1–0.5 range. Tuna fish with its fairly large r value has a larger proportion of medium users than most categories. This yields less bowed concentration curves.

NUNBD Excluding the Zeroes

In this section we consider the observed actual concentration *excluding* those who made zero purchases. The estimated long-run concentration is again based on (4), but the "$g(\lambda)$" and "$E[\lambda]$" are based only on the 88 percent of the consumers with $\lambda > 0$. (To be more precise, the probability density and the expectation used in (4) would be $g(\lambda \mid \lambda > 0)$ and $E[\lambda \mid \lambda > 0]$, respectively.) Table 4 gives the results.

TABLE 4
Concentration among Users, I.E., Excluding "Zeroes"
NUNBD *Model Versus Observed*
Tuna Purchases

Cumulative % Households	10	20	25	30	40	50
1-Month Actual % Purchases	22.49	37.33	43.72	50.10	61.62	68.01
1-Month Model % Purchases	23.59	38.61	44.86	51.10	62.54	68.78
3-Month Actual % Purchases	27.33	44.42	51.41	57.41	67.90	76.49
3-Month Model % Purchases	27.35	44.44	51.43	57.73	68.23	76.52
6-Month Model % Purchases	28.57	46.39	53.65	59.90	70.72	79.25
Estimated Long-run % Purchases	29.21	47.23	54.69	61.09	71.95	80.67

NUNBD Parameters: $r = 1.383$; $\alpha = 0.677$; $q = 0.123$.

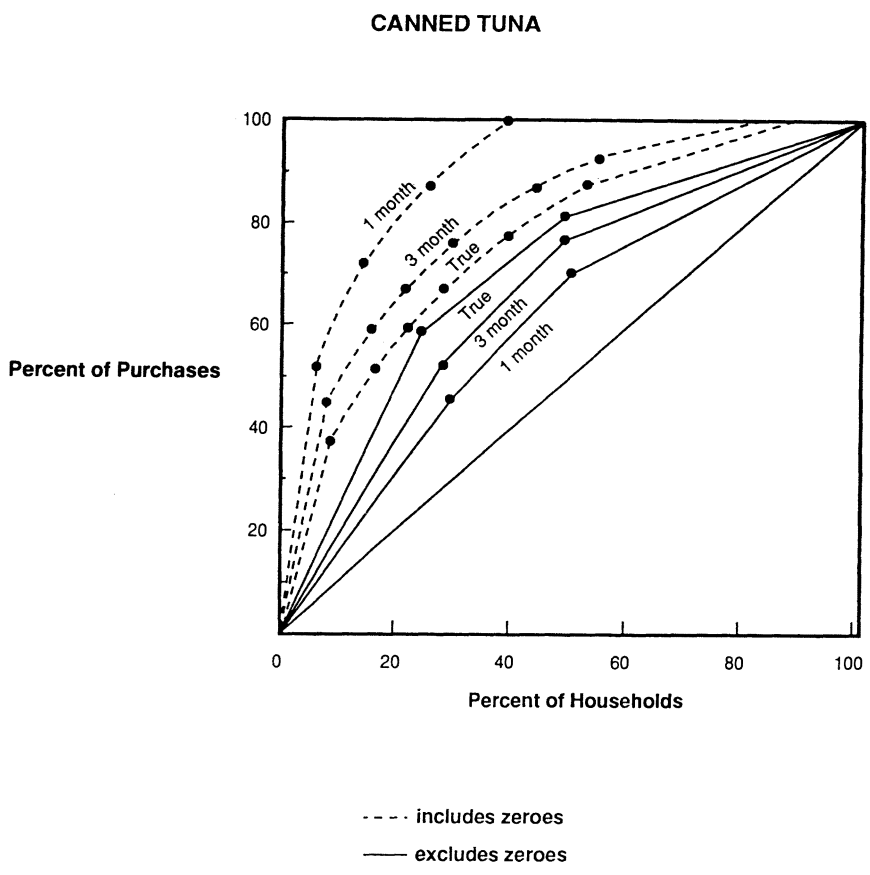


FIGURE 2. Sources of Variation in Apparent Customer Concentration Levels

First we see that the effect of time is opposite to that in Table 3. As time *increases*, the observed concentrations *increase* to the true concentration. (We saw the same phenomenon for catsup and yogurt purchases in Table 1.) This happens because we are excluding zero purchasers. In one month those who purchase only have time to buy once or twice. Thus, *among users* there is very little concentration. As time increases, the heavy users can buy five or six times and the light users can make their single purchase. Thus, among users the concentration curve becomes more bowed as time increases.

Finally, the estimated long-run concentrations are slightly different in Table 3 than in Table 4. This happens because Table 3 reports true concentration among all households, and Table 4 reports it only among the proportion $1 - q = 0.88$ of households who are users of the product. Since the proportion of hard core nonusers, q, is not too large here (only 0.12), the two true concentration curves are fairly similar.

Typical Results

Figure 2 graphs the results presented in Tables 3 and 4.[3] Recall that in Table 3 we include the nonusers (apparent, or true, as indicated in the table), and with the NUNBD model we exclude the nonuser class. Figure 2 shows very clearly that:

[3] In the figure we have plotted several points on each Lorenz curve, and connected those points with a straight line. The Lorenz curves themselves would of course be smoother than the lines drawn. The points plotted simply serve to illustrate the main qualitative characteristics noted below.

- Observed concentration among the entire population of users and nonusers generally *decreases* to truth as time increases.
- Observed concentration among users generally *increases* to truth as time increases.
- True concentration in the entire population is estimated to be higher than true concentration among users only. However, the next data set will show that the second of these statements does not always hold in every product category.

Toilet Tissue

Table 5 is in the same format as Tables 3 and 4. The NUNBD model is used with the resulting parameters: $r = 4.532$, $\alpha = 1.241$, $q = 0.040$. Note the small values for nonusers, $q = 0.04$. (The reader should not worry—this value is due to toilet paper bought in outlets not covered by the panel.) Thus, even though Table 5 uses the NUNBD model, the small q-value makes it almost the NBD model. The excluded observed zeroes make the observed concentration *increase* from one to three months. The "close to NBD" spirit of the $q = 0.04$ NUNBD model makes the observed concentrations *decrease* from three to six months. The "typical" monotonicity of Tables 3 and 4 does not hold for the toilet-tissue data.

Product Categories with Low r-Values

Table 6 shows the results for frozen orange juice. The NUNBD parameters are: $r = 0.182$, $\alpha = 0.144$, $q = 0.067$. The low q value means that most panelists eventually purchase. However, the low r value means that consumers vary greatly with respect to their true purchasing rates. (A gamma distribution with $r = 0.182$ is a *highly* skewed reverse J-shaped distribution.) This low r value makes both the observed and true concentration curves very bowed—or highly concentrated. Table 6 shows the typical monotonically increasing pattern of observed concentrations. Table 7, which examines concentration in the entire population (i.e., users and nonusers) shows dramatically concentrated (bowed) inverted Lorenz curves that decrease monotonically towards the true curve.

Concentration at the Brand Level

All of our previous results are based on product-category purchases. We can do analogous analyses at the brand level. Table 8 contains the NUNBD approach to Scott Toilet Tissue. That is, the across-consumer three-month histogram of only Scott purchases is

TABLE 5

Concentration among Users
NUNBD *Model Versus Observed*
Toilet Tissue Purchases

Cumulative % Households	10	20	25	30	40	50
1-Month Actual % Purchases	22.62	38.85	44.42	49.84	60.68	71.52
1-Month Model % Purchases	22.99	39.25	45.51	50.93	61.77	72.61
3-Month Actual % Purchases	N.A.	40.85	47.81	54.08	64.79	74.24
3-Month Model % Purchases	23.26	39.64	46.61	52.99	64.15	74.05
6-Month Model % Purchases	22.14	38.13	45.02	51.35	61.73	72.72
Estimated Long-Run % Purchases	19.61	34.50	41.07	47.21	58.34	68.15

NUNBD Parameters: $r = 4.532$; $\alpha = 1.241$; $q = 0.040$.

TABLE 6

Concentration among Users
NUNBD *Model Versus Observed*
Orange Juice Purchases

Cumulative % Households	10	20	25	30	40	50
1-Month Actual % Purchases	24.23	41.96	50.26	55.59	65.35	75.11
1-Month Model % Purchases	30.42	46.89	53.87	59.14	68.44	76.74
3-Month Actual % Purchases	*	*	*	*	*	*
3-Month Model % Purchases	37.14	55.56	62.32	68.04	76.97	83.77
6-Month Model % Purchases	41.05	60.30	67.12	72.77	81.36	87.44
Estimated Long-Run % Purchases	65.99	86.02	90.95	94.42	97.99	99.38

* Not available due to group (5^+ category) in histogram.
NUNBD Parameters: $r = 0.182$; $\alpha = 0.144$; $q = 0.067$.

used to estimate the NUNBD parameters. Comparing the Scott brand to the category as a whole is illuminating:

	Scott	Category
$r =$	0.832	4.532
$\alpha =$	0.443	1.241
$q =$	0.774	0.040

The category is *very* homogeneous in purchase rates ($r > 4$) while Scott is quite heterogeneous ($r < 1$). The product category has virtually no nonusers while Scott has only about one quarter of the consumers buying its brand. (The α parameter is merely a scaling parameter which by itself has no meaning.) Thus, a brand can behave *very* differently from the category. We will return to this point in the last section on strategic implications.

Summary

The results from the tuna fish, toilet paper and frozen orange juice categories show clearly that:

- *Including* the zero purchasers in the observed concentration curves usually *overstates* true concentration.
- *Excluding* the zero purchases typically *understates* true concentration.
- The above two effects usually *decrease* as the time of observation *increases*.

TABLE 7

Concentration among All Households
NUNBD *Model Versus Observed*
Orange Juice Purchases

Cumulative % Households	10	20	25	30	40	50
1-Month Actual % Purchases	78.96	100	100	100	100	100
1-Month Model % Purchases	78.93	100	100	100	100	100
3-Month Actual % Purchases	70.36	91.40	95.96	100	100	100
3-Month Model % Purchases	72.13	92.10	96.34	100	100	100
6-Month Model % Purchases	70.19	89.84	94.59	97.36	100	100
Estimated Long-Run % Purchases	68.30	87.54	92.36	95.46	98.54	99.63

NUNBD Parameters: $r = 0.182$; $\alpha = 0.144$; $q = 0.067$.

TABLE 8

Concentration among Users
NUNBD *Model Versus Observed*
Scott Toilet Tissue Purchases

Cumulative % Households	10	20	25	30	40	50
1-Month Actual % Purchases	23.83	40.65	46.51	52.12	63.34	71.89
1-Month Model % Purchases	24.77	40.45	46.41	52.37	64.25	70.21
3-Month Actual % Purchases	29.76	47.71	54.59	60.89	70.76	78.72
3-Month Model % Purchases	29.54	47.24	53.90	60.32	70.32	78.18
6-Month Model % Purchases	31.67	50.04	57.24	63.42	73.79	81.76
Estimated Long-Run % Purchases	35.79	55.37	62.88	69.30	79.37	86.93

NUNBD Parameters: $r = 0.832$; $\alpha = 0.443$; $q = 0.774$.

- The above three statements are not *always* correct; see the toilet-tissue data in Table 5.
- It takes a small r value (i.e., large variability across households in true purchasing rates) for true concentrations to be as high as those implied by the 80/20 "law"; see the orange-juice data in Tables 6 and 7.
- A modeling approach that allows for consumers who will never buy the product, e.g., the NUNBD, is preferred. The nonusers can be removed from the analysis and the resulting concentration curves will apply to the users only.

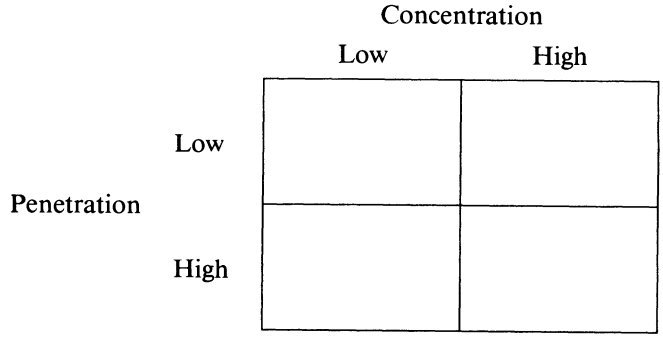
For the seven products examined in this paper, one-year concentration statistics tended to be fairly close to the true long-run concentration levels (while shorter time periods for observation did not meet this standard). The one exception occurred for yogurt purchases, which even after a year had not settled at the long-run concentration level. Of course, one year will not always be "long enough" for products bought either less frequently or less regularly (including, for example, more durable products and many services). That is, for such products the observed differences in customer purchase histories will be dominated by the unstable within-customer variation rather than the long-run differences across customers in purchase rates. The more important conclusions are that the modelling approach allows one to:

(1) anticipate the kinds of products for which discrepancies between observed and long-run concentration will be relatively severe, and

(2) adjust observed concentration statistics to better reflect the source (i.e., long run stable purchase patterns versus transitory, random influences) of that observed concentration in customer purchasing.

Strategic Implications

Penetration and concentration are useful criteria for strategically classifying markets. Consider, for example, the 2 × 2 classification below:



In low-penetration markets, in general, the need to increase awareness and trial is obvious. This is particularly important for low-penetration, low-concentration markets. In high-concentration markets, firms often battle over the loyalty of the heaviest users, but must ask if this is being done at the sacrifice of profitability. Particularly for high-penetration, high-concentration markets, in battling over market shares, price competition can lead to an increasing spiral of price promotions without a great deal of regard for margins. In low-penetration, high-concentration markets, firms must ask if they are purposefully pursuing a niche strategy or wish to approach a mass market. High-penetration, low-concentration markets may require extensive distribution and (for durables) extensive service support.

At this point the reader may disagree with some of our brief analysis of the above four quadrants. However, one must admit that concentration and penetration are two very important dimensions whenever strategic issues are discussed. Given this, the NUNBD model changes from a "toy" for the probability modeler into a "tool" for the marketing manager. The two most relevant parameters are r, which measures the proportion of buying rates directly, and q, the proportion of hard-core nonbuyers. Recall that since

$$\frac{\sqrt{\text{Var}\,[\lambda]}}{E[\lambda]} = \frac{1}{\sqrt{r}}$$

we see "concentration" *decreases* as r *increases*. The proportion of nonusers q is, of course, exactly the complement of "penetration."

The NUNBD parameters can be used directly to put brands or product categories into the four cells of the 2×2 concentration/penetration matrix. In fact, since these two key parameters, r and q, are continuous, we could more accurately position each brand or product category in the two-dimensional $r - q$ plane.

Since the primary purpose of this paper is methodological, we will not discuss further the strategy implications per se for any of the specific products discussed. We also do not wish to oversell our results. Clearly other data and additional dimensions will come into play. Therefore, our limited—but we feel important—contributions to marketing strategy are as follows:

- Concentration and penetration are important dimensions for marketing decisions.
- Observed concentration and the observed proportion of buyers can be very badly biased estimates of the true underlying constructs.
- A formal model is needed to estimate "truth" from observation.
- The NUNBD model has an appealing mathematical "story" and its parameters are direct measures of concentration and penetration.

The reader who has made it this far should now have a healthy skepticism regarding any 80/20 "laws." In any event, that old marketing tautology, "There are two ways to get more business: (1) get more customers, and (2) get more business from existing customers," still holds. In searching for truth in concentration, our NUNBD framework focuses attention on these two points. It also eliminates the ambiguity in examining observed concentration among households that happen to have purchased during a single time period of a certain duration.

Acknowledgements. The authors thank NPD and A. C. Nielsen as firms providing the customer purchase information analyzed in this paper.

This paper was received July 2, 1991, and has been with the authors 4 months for 3 revisions. Processed by Scott A. Neslin, Area Editor.

Appendix. The NUNBD Model

This model assumes that some proportion q of the population are hard-core nonusers, i.e., have a purchase rate $\lambda = 0$. The remaining proportion $1 - q$ follows the usual NBD model, with Poisson purchasing, and purchase rates distributed gamma across these users as in Equation (2) in the text.

The aggregate purchase-frequency distribution for the combined population of users and nonusers is therefore

$$P_{\text{NUNBD}}(X = 0) = q + (1 - q)P_{\text{NBD}}(X = 0) \quad \text{and} \tag{A1}$$

$$P_{\text{NUNBD}}(X = x) = (1 - q)P_{\text{NBD}}(X = x) \quad \text{for } X > 0, \tag{A2}$$

where $P_{\text{NBD}}(X = x)$ is given in equation (2). In addition to the parameter q, representing the proportion of hard core nonbuyers, the remaining two NUNBD parameters are also easy to interpret. r/α is the mean purchase rate among users, and r, the shape parameter of the gamma mixing distribution, is an index of homogeneity in purchase rates across users. That is, a large value of r indicates that purchase rates do not vary greatly across users, and a small r-value suggests the reverse. (To see this, note that $r^{-1/2}$ is the coefficient of variation of the gamma distribution for purchase rates across users.)

Explicit formulas for maximum-likelihood estimates of r, α, and q cannot be obtained, so an iterative pattern search is employed to compute the MLEs used in this paper.

References

Chatfield, C. and G. J. Goodhardt (1973), "A Consumer Purchasing Model with Erlang Inter-purchase Times," *Journal of the American Statistical Association*, 68, 828–835.

Cowell, F. A. (1977), *Measuring Inequality*, New York: Wiley.

Curry, B. and K. D. George (1983), "Industrial Concentration: A Survey," *The Journal of Industrial Economics*, 31, 203–255.

Ehrenberg, A. S. C. (1988), *Repeat Buying*, 2nd Ed., New York: Oxford University Press.

Gupta, S. (1988), "Impact of Sales Promotions on When, What, and How Much to Buy," *Journal of Marketing Research*, 25, 342–356.

Herniter, J. (1971), "A Probabilistic Market Model of Purchase Timing and Brand Selection," *Management Science*, 18, Part II, 102–113.

Kahn, B. E. and D. G. Morrison (1989), "A Note on 'Random' Purchasing: Additional Insights from Dunn, Reader and Wrigley," *Applied Statistics*, 38, 111–114.

———, ——— and G. Wright (1986), "Aggregating Individual Purchases to the Household Level," *Marketing Science*, 5, 260–268.

——— and D. C. Schmittlein (1989), "Shopping Trip Behavior: An Empirical Investigation," *Marketing Letters*, 1, 55–70.

Lawrence, R. J. (1980), "The Lognormal Distribution of Buying Frequency Rates," *Journal of Marketing Research*, 17, 212–220.

Lord, F. and M. R. Novick (1968), *Statistical Theories of Mental Test Scores*, Reading, MA: Addison-Wesley.

Morrison, D. G. (1969), "Conditional Trend Analysis: A Model That Allows for Nonusers," *Journal of Marketing Research*, 6, 342–346.

——— and D. C. Schmittlein (1981), "Predicting Future Random Events Based on Past Performance," *Management Science*, 27, 1006–1023.

——— and ——— (1988), "Generalizing the NBD Model for Customer Purchases: What Are the Implications and Is It Worth the Effort?" *Journal of Business and Economic Statistics*, 6, 145–160.

Schmittlein, D. C. (1988), "Issues in Measuring Market Concentration among Firms, Suppliers, and Customers," Working Paper, Marketing Department, The Wharton School, University of Pennsylvania.

——— and D. G. Morrison (1983), "Prediction of Future Random Events with the Condensed Negative Binomial Distribution," *Journal of the American Statistical Association*, 78, 449–456.

Twedt, D. W. (1964), "How Important to Marketing Strategy Is the 'Heavy User'?" *Journal of Marketing*, 28 (January), 71–72.

Wheat, R. D. and D. G. Morrison (1990), "Estimating Purchase Regularity with Two Interpurchase Times," *Journal of Marketing Research*, 27, 87–93.