



*The World's Largest Open Access Agricultural & Applied Economics Digital Library*

**This document is discoverable and free to researchers across the globe due to the work of AgEcon Search.**

**Help ensure our sustainability.**

Give to AgEcon Search

AgEcon Search

<http://ageconsearch.umn.edu>

[aesearch@umn.edu](mailto:aesearch@umn.edu)

*Papers downloaded from **AgEcon Search** may be used for non-commercial purposes and personal study only. No other use, including posting to another Internet site, is permitted without permission from the copyright owner (not AgEcon Search), or as allowed under the provisions of Fair Use, U.S. Copyright Act, Title 17 U.S.C.*

*No endorsement of AgEcon Search or its fundraising activities by the author(s) of the following work or their employer(s) is intended or implied.*









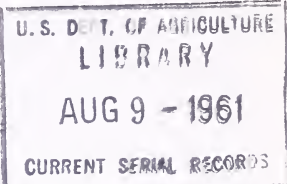
## Historic, archived document

Do not assume content reflects current scientific knowledge, policies, or practices.

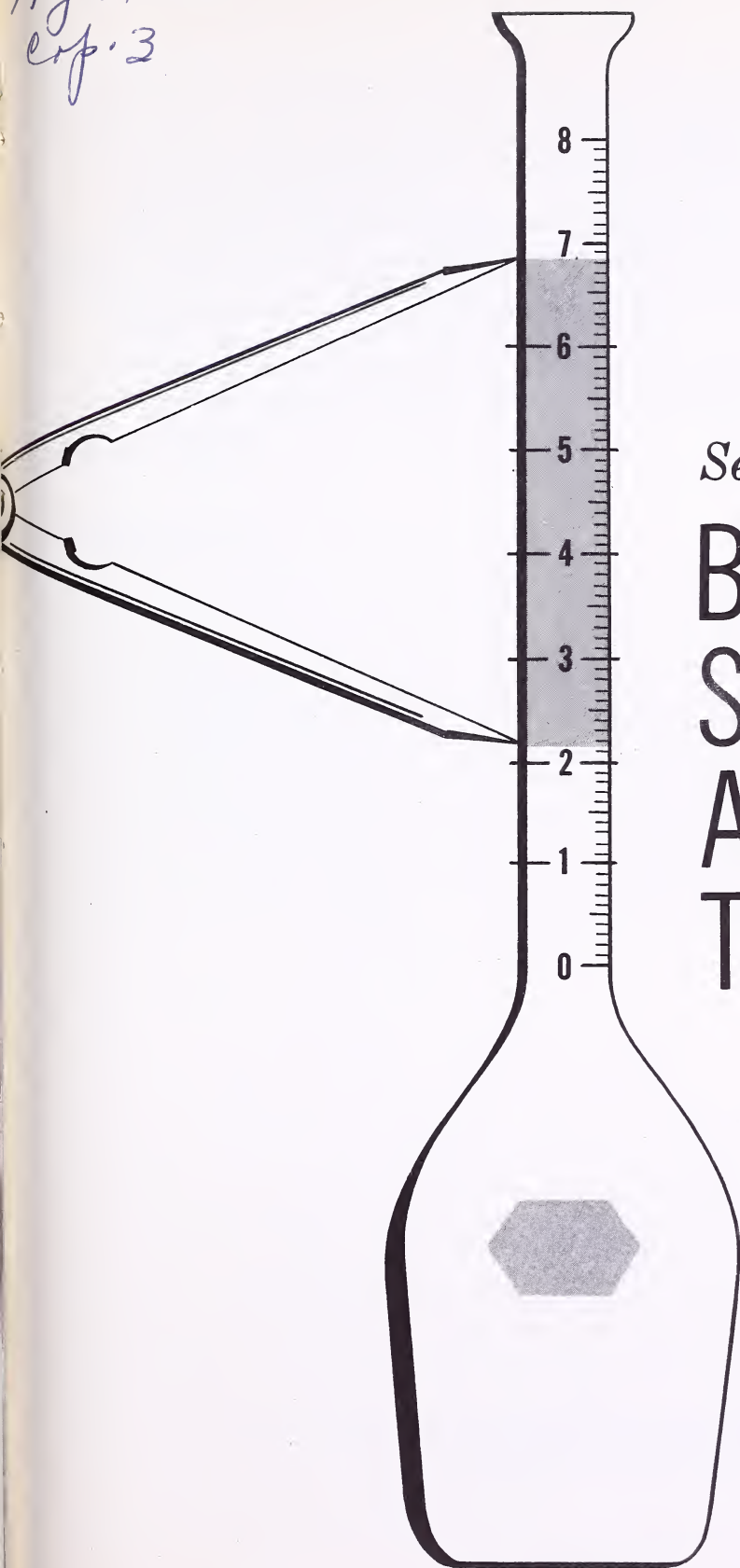


1  
Ag 845m  
exp. 3

Marketing Research Report No. 482



*Selected Problems In*  
**BUTTERFAT  
SAMPLING  
AND  
TESTING**





## CONTENTS

	<u>Page</u>
Summary and conclusions . . . . .	4
Introduction . . . . .	6
Problems with application for all butterfat sampling and testing . . . . .	7
Differences among testers . . . . .	7
Pipetting temperatures . . . . .	10
Storing and reheating composite samples . . . . .	14
The blending of can milk in plant weigh tanks . . . . .	19
Problems related to pick up of milk in bulk tanks . . . . .	21
Sampling from bulk tank trucks . . . . .	22
Adding preservatives to the farm sample . . . . .	25
Transporting samples from the farm to the laboratory . . . . .	26
Including defective daily portions in building composite samples . . . . .	28
Literature cited . . . . .	33

Washington, D. C.

June 1961



## SUMMARY AND CONCLUSIONS

Experiments were conducted on three problems having general application to Babcock butterfat test results, on one problem affecting only tests on milk received in cans, and on four problems related to milk picked up in bulk tanks. The results provide a basis for increased confidence in methods of testing and of verifying tests of milk delivered to plants by milk producers. Highlights of the results are as follows:

1. Variations among testers in measuring and reading the Babcock test frequently result in test differences of one point on identical samples by two experienced testers, but seldom result in differences greater than one point. Indications are that over 98 percent of the tests on the same samples by pairs of testers can be expected to be within one point of each other. Such agreement is found among testers who regularly test large numbers of samples and among testers who frequently compare results and methodology or have the same supervision. Technicians who test only occasionally or who work under different supervision can be expected to have test results that do not agree this closely.

The estimated part of the test variance which could be eliminated by using only one pipetter ranged from 0 to 29 percent, and the part of the variance which could be expected to be eliminated by using only one reader ranged from 0 to 41 percent.

2. In only one of five markets where the relation between pipetting temperatures and test results was studied, was there a consistent and significant relationship between pipetting temperatures and test results, with higher temperature yielding lower tests.

3. The number of times a composite sample was reheated was a more important factor affecting level of test than the length of time the composite was stored. The samples differing the most from the original composite tests were those held 5 days and reheated three times. On the average, they tested 0.13 lower in percent of butterfat than the corresponding original composite samples.

4. In plants receiving milk in cans, when some method of agitation was used to improve the mixing of the milk in the weigh tanks before sampling, from 0 to 36 percent of the samples differed in test by more than 0.1 in percent (or 1 point) of butterfat. Even among tanks with the same method of agitation, there was considerable variation in the percentages of paired samples differing by more than 1 point. In 5 plants where weigh tanks were sampled for experimental purposes, without previously agitating the milk, from 10 to 50 percent of the paired samples differed by more than 1 point. Where the samples were taken following some form of agitation, none of the average differences differed significantly from zero.

5. Differences between the weighted averages of producers' tests and the test of milk taken from loaded tank trucks of milk varied considerably according to the method of sampling from the tank truck. The experiments indicated that representative samples can be obtained from loaded tank trucks only if the milk has been agitated before sampling. In the experiments where some form of agitation was used before sampling the tank, 97 percent or more of the samples from the tank truck tested within 0.04 percent of butterfat of the weighted average of the tests on the individual herd milks commingled in the tank truck.

6. The use of four drops of preservative in the sample bottle before pipetting

was found to have no significant effect on the tests of fresh samples.

7. Test results on composite samples built and pipetted at the farm were compared with tests on composites built and pipetted at the laboratory, and tests on fresh samples pipetted at the farm were compared with tests on fresh samples pipetted at the laboratory. Results on samples pipetted at the farm did not differ significantly from tests of samples pipetted at the laboratory. It was concluded, therefore, that transportation of samples on the tank truck did not significantly affect the samples. Usual precautions taken to preserve samples of bulk milk between sampling at the farm and testing at the laboratory appear to be adequate. Exceptions can occur, however, through carelessness of individual haulers, particularly on warm days. The effect of such mistreatment of samples was not determined in this study.

8. When, for experimental purposes, some defective daily portions, such as frozen or churned, were included in the composite sample, the spread between the tests of the composite and the averages of the tests of the daily fresh samples was greater in a significant number of cases than when all daily portions in the composite were normal.

# SELECTED PROBLEMS IN BUTTERFAT SAMPLING AND TESTING

By Anthony G. Mathis, Robert W. Johnson, and  
Elsie D. Anderson 1/  
Marketing Economics Division  
Economic Research Service

## INTRODUCTION

This is a report of eight studies of techniques in sampling milk and testing it for butterfat content.

Most of the methods studied have been suggested by Federal order market administrators or others as subjects needing research, because there was insufficient reliable knowledge as to whether they cause significant difference in test results.

If duplicate series of samples are tested by techniques differing even slightly, the average percentages of butterfat and the variations around these averages may be different.

To lessen the impact of bias and test variability on producers' returns for milk, Federal market administrators, producers' organizations, and some State agencies check the butterfat tests performed by plants. Market administrators also test fluid milk products of each plant to verify the plant's report of milk usage. Plants buying milk from other plants check the seller's statement of butterfat content.

As an administrative necessity, an official test is regarded as correct, and undue deviation of the plant tests from this norm is subject to correction. In the case of two plants, one buying, the other selling milk, differences in the plants' test results are a subject for negotiation before settling for the milk. As a practical matter, the comparison between the tests must be defined in terms of a range about the check-test within which a plant's test is acceptable. This range of acceptability is necessary in part because the Babcock test which is the accepted test for butterfat in the United States, is accurate only within bounds. One of the factors limiting the accuracy of the Babcock test is inaccuracy in the calibration of glassware used in the test (4, 5). 2/ The Association of Official Agricultural Chemists makes very specific recommendations as to the type of glassware that should be used. However, a number of States have no specifications for glassware (9). 3/

Variations in test results, beyond those inherent in the Babcock test, also may be caused by a number of factors, such as slight differences between individuals in performing testing routines, difference in the representativeness of samples, and

---

1/ Fred Stein, formerly with Marketing Economics Division, and now with the Dairy Division, Agricultural Marketing Service, had an active part in planning this work and made arrangements with market administrators and other sources of original material for the collection of data for this report.

2/ Underscored figures in parentheses refer to items in Literature Cited, page 33.

3/ In order to minimize differences caused by glassware, all glassware used in the present study was officially calibrated by State Agricultural Colleges or qualified laboratories.



slight differences in procedures. Variations of this kind can be controlled, and to this end the present study is addressed. Specifically, the task of this present study is:

1. To determine how and to what extent the use of different techniques in sampling and testing may affect test results.

2. To indicate biases and variations in test results along with probability that variations of a given size will occur when a given technique is used.

These objectives should afford persons or organizations interested in butterfat tests a broader base of empirical knowledge for deciding the acceptability of various procedures and for establishing limits within which tests on the same sets of samples can be expected to agree.

Most of the studies included in this report involved closely controlled procedures. Comparisons, therefore, usually were made on a limited number of samples. The studies have been grouped into general problems in butterfat sampling and testing, a sampling problem for plants receiving milk in cans, and problems related to milk picked up in bulk tank trucks. In general, analysis of variance was used for statistical comparisons of results from different methods of sampling or testing. The number of observations, plants, and markets differed from one study to another. Also, the statistical methods used in the analysis differed among studies. The methods followed will be identified in the discussion of each study. Generally, the findings are expressed by giving the proportions of test results that can be expected to agree or differ by stated amounts.

These studies were concurrent with a larger research project in which about 230,000 milk samples were tested from deliveries of 1,700 producers for an average of 5.5 months to 21 "plants" in 9 markets. <sup>4/</sup> The principal objective of the larger project was to determine how much variability in producers' tests can be expected under certain environmental conditions, whether this expected variability is constant from season to season and market to market, and how butterfat sampling and testing programs can be organized to take into account this normal variability.

## PROBLEMS WITH APPLICATION FOR ALL BUTTERFAT SAMPLING AND TESTING

Three of the problems considered in this study have general application to Babcock test results. These problems pertain to the actual testing of milk in the laboratory: differences among testers and departures from standard methods, specifically pipetting temperatures and techniques affecting tests on composite samples.

### Differences Among Testers

The procedures involved in the Babcock test are carefully defined to limit the possibilities for differences in test results on the same sample among different testers or by the same tester in repeated testing. However, there remain possibilities for differences in tests because of varying personal abilities of the individual testers to make measurements and read tests uniformly.

---

<sup>4/</sup> The can and bulk operations in each of 2 plants were considered to be separate operations, so that the 21 "plants" represented 19 establishments.

Results of experiments carried out by Herreid and others indicate that differences in test results can be lessened, in many cases, by more careful supervision and attention to techniques (5, 6, 7, 8). Herreid points out that many testers fill their pipettes to the point where the lowest part of the meniscus is level with the mark, although Babcock considered that the pipette was full when the milk touched the mark. (5).

A number of studies show that the use of glymol to eliminate the meniscus would lessen variability in reading tests. Herreid found that increasing the size of sample from 18 to 18.36 grams as well as using glymol would bring the Babcock method into closer agreement with the Rose-Gottlieb method. Lampert, Nelson, and Wilster (13) and Herreid (5) suggest the use of reading devices that would improve accuracy.

A recent study involving tests of duplicate samples by six technicians in different laboratories affords some measure of the ability of testers to reproduce Babcock test results (11). This study, in which the tests were read to the nearest one-hundredth of one percent, showed that in two-thirds of replicated tests, the same tester would get results from one test within 0.046 percent butterfat of another test on the same sample. From this figure one can deduce that 95 percent of paired comparisons of tests which have been read to the nearest 0.01 percent would be within 0.092 percent, and 99 percent would be within 0.138 percent. The standard deviation of the difference between two readings which would be expected due to the rounding of test readings to the nearest 0.01 percent would be  $\pm 0.0047$ . It appears that rounding was a minor factor in the differences between readings for these comparisons.

Babcock tests are almost always read to the nearest one-tenth of one percent. Each reading involves a maximum error of  $\pm 0.05$ , due to rounding. If the tests being rounded are distributed uniformly over the 0.1 percent interval, or from .05 below to .05 above the rounded reading, two-thirds of the tests would be included in the interval from .033 below to .033 above the rounded reading. Based on this estimated standard deviation of 0.033 for the rounding of individual test readings, the standard error of an average of 30 daily tests, due to rounding, would be  $\pm .006$ , and the standard deviation of the difference between two tests which might be attributable to the rounding procedure would be  $\pm .047$ .

The question of a tester's personal bias in the reading of a test must also be considered. A small amount of bias will not affect the reading of each individual test, although it affects the average reading of a group of tests by the amount of the bias when the usual rounding procedures are followed, if the tests are evenly distributed over the rounding interval. With a 0.01 percent bias, an average of 1 in every 10 tests would be expected to differ by 0.1 percent butterfat (or 1 point) after rounding to the nearest 0.1 percent. The average of the 10 tests would then be changed by 0.01 percent or by the amount of the bias. A bias of 0.02 percent would be reflected by a 0.1 percent (or 1 point) difference on 2 in every 10 readings. The average of the 10 tests would then be changed by 0.02 percent or by the amount of the bias.

The effect of a bias on butterfat test readings and averages can be verified by starting with a series of 10 true readings such as 4.00, 4.01, to 4.09, rounding them, and comparing the rounded percentages and their average with the rounded percentages and average of a series in which the same true readings have had the bias added (or subtracted) before rounding and averaging. For example, adding a 0.01 percent bias to the example above would change the series to 4.01, 4.02 to 4.10 and would increase the average of the rounded percentages from 4.04 to 4.05.

Information about the numbers of differences of a given size is more important



than average differences. Such frequency distributions afford a basis for deciding when the disagreement between two sets of results covering the same producers' milk is within bounds that may normally be expected.

Test results tend to differ more among technicians who test occasionally than among those who regularly test large numbers of samples. Also, results appear to vary more among testers who do not regularly work together, or who work under different supervision, than among testers who frequently compare results and methodology or have the same supervision (table 1). In one report it was suggested that "...psychology may influence a test. A majority of testers are subject to influence and suggestion...Testers working under too critical scrutiny may readily be influenced by the attitude of employers...." 5/

In 7,192 comparisons of paired results on identical samples, by testers accustomed to working together, 74 percent agreed, 25 percent differed by 1 point, and 1 percent differed by 2 points or more (table 1). In the 6 experiments with 4 testers in each, tests agreed on 69.4 to 81.9 percent of the paired samples (table 2). These comparisons indicate rather clearly that variations in measuring and reading the Babcock test frequently result in test differences of one point on identical samples by two experienced testers, but seldom result in differences exceeding one point. More than occasional differences larger than one point warrant an examination of the sampling or testing procedures used by the testers.

Part of the differences among testers' results on duplicate samples is due to small variations in techniques. Given uniform techniques, some differences can be

Table 1.--Percentage of duplicate samples of milk given same and different Babcock readings by two testers, by working relationship of testers

Description of testers	Pairs of duplicate samples	Pairs of tests in agreement	Pairs of tests differing by:		
			1 point	2 points	Over 2 points
	Number	Percent	Percent	Percent	Percent
Testers working closely together:					
Tester and check-tester,					
1 market 1/ .....	6,760	74.2	24.6	1.0	0.2
Market administrators' testers,					
6 markets 2/ .....	432	76.9	22.9	.2	0.0
Total .....	7,192	74.4	24.5	.9	.2
Testers not accustomed to working together:					
31 technicians from various					
plants 3/ .....	1,846	54.1	36.2	7.9	1.8
8 research technicians 4/ .....	1,834	26.9	42.9	19.2	11.0
Total .....	3,680	40.6	39.5	13.5	6.4

1/ Duplicate samples of milk taken from a storage tank at the same time by tester and check-tester. Most of the testing was done by two men.

2/ 6 experiments, 4 testers in each, using subsamples of 12 samples.

3/ Chiefly plant testers at refresher course at the University of Minnesota, using subsamples of 4 samples.

4/ Technicians who did occasional testing; each from a different agency.

5/ Unpublished report of a refresher course for butterfat testers held at the University of Minnesota, Feb. 1958.

Table 2.--Percentage of duplicate samples of milk given same and different Babcock readings by two testers, 6 experiments

Experiment	Pairs of duplicate tests <u>1/</u>	Pairs of tests in agreement <u>2/</u>	Pairs of tests differing by:		
			1 point	2 points	Over 2 points
	Number	Percent	Percent	Percent	Percent
Number 1 .....	72	81.9	18.1	0	0
Number 2 .....	72	81.9	18.1	0	0
Number 3 .....	72	77.8	22.2	0	0
Number 4 .....	72	76.4	23.6	0	0
Number 5 .....	72	73.6	25.0	1.4	0
Number 6 .....	72	69.4	30.6	0	0
Total .....	432	76.9	22.9	.2	0

1/ In each experiment, 4 operators tested 12 samples; this affords 72 comparisons of readings by two testers.

2/ Tests differing by from  $-.05$  to  $+.05$  were considered to be in agreement.

controlled by improving testers' skills and care. In the present study several experiments were made to determine how much of the variation was caused by differences in pipetting the sample into the test bottle and by differences among testers in reading the completed test.

In each of five markets tests on duplicate samples were prepared by two pipettors and read independently by two readers. In a sixth market, tests on samples in 6 experiments were prepared by one pipettor and read independently by from 5 to 15 readers.

Differences between testers in pipetting samples caused highly significant differences in testers' results in three of five experiments. In the other two experiments, differences were too small to be significant (table 3, markets 1, 2, 5, 6, 8). In these five experiments, differences among testers in reading tests were highly significant in only one trial. In each of the 6 experiments in which one man pipetted and prepared tests and several technicians read each test result, differences among the individual readers were highly significant (table 3, Market 9).

The estimated amount of test variance which could be eliminated by using only one pipettor ranged from 0 to 29 percent. Three out of five of the experiments produced estimates of 19 percent or less. That part of the variance which could be expected to be eliminated by using only one reader ranged from 0 to 41 percent. Seven of 11 experiments yielded estimates of 21 percent or less (table 4).

#### Pipetting Temperatures

It has been a common practice for testers to pipette fresh samples at  $70^{\circ}$  F. and composite samples at  $100^{\circ}$  F. despite the fact that the Association of Official Agricultural Chemists specifies  $100^{\circ}$  F. as the standard pipetting temperature for both kinds of samples (12). In fact, pipetting temperatures recommended in State regulations have varied widely (table 5).

Previous work has shown that the tests of fresh samples give results significantly higher than those of composite samples (10, 17). Theoretically, a high pipetting temperature could cause a lower test than a low pipetting temperature,

Table 3.--Effect on test results of difference among technicians in pipetting and reading Babcock tests of milk samples

			: Average test results : : on samples : : pipetted by-- :			: Level of significance : : of differences in : : average tests on : : samples grouped by--3/ :			Components : of variance, : estimated : standard error		
Market: Samples: Pipetters: Readers:			: Lowest : Highest : : average : average : : test : test :			: Pipetters : Readers : : averages : averages : : 4/ : 5/					
: Number	: Number	: Number	Fat %	Fat %	Fat %	Percent	Percent	Fat %	Fat %	Fat %	
1 .....	12	2	4.42	4.43	4.41	none	none	0	0.024	0.024	
2 .....	12	2	4.76	4.79	4.78	1	none	.024	0	0	
5 .....	12	2	4.57	4.60	4.57	1	1	.019	.028	.028	
6 .....	12	2	3.90	3.95	3.91	none	none	.032	.013	.013	
8 .....	12	2	4.07	4.11	4.09	1	none	.031	.012	.012	
9 .....	36	1	4.61	--	4.59	--	1	--	.016	.016	
9 .....	36	1	4.30	--	4.26	--	1	--	.025	.025	
9 .....	36	1	4.65	--	4.59	--	1	--	.027	.027	
9 .....	36	1	4.34	--	4.31	--	1	--	.017	.017	
9 .....	36	1	4.08	--	4.06	--	1	--	.013	.013	
9 .....	36	1	4.05	--	4.01	--	1	--	.023	.023	

1/ The averages for markets 1, 2, 5, 6, and 8 are the averages of the tests on all the samples pipetted by each of the two pipetters. One man pipetted all of the samples in each experiment in market 9, and the averages for market 9 are the averages for all of these tests.

2/ The averages for markets 1, 2, 5, 6, and 8 are the averages of the tests on all the samples read by each of the two readers. The averages for market 9 were the average for the reader having the lowest average and for the reader having the highest average test.

3/ Significance of the differences tested by analysis of variance.  
4/ The average test on samples pipetted by either of the pipetters could be expected, two out of three times, to fall within plus or minus one standard error of the average test for samples pipetted by both pipetters.

5/ The average test on samples read by one reader could be expected, two out of three times, to fall within plus or minus one standard error of the average test as read by all readers.



Table 4.--Variance in test results due to differences in pipetting milk samples and in reading the Babcock test

Item	2 readers, 2 pipettors, 12 samples each in--					
	Market 1	Market 2	Market 5	Market 6	Market 8	
	<u>Fat %</u>	<u>Fat %</u>	<u>Fat %</u>	<u>Fat %</u>	<u>Fat %</u>	
Within-sample variance due to:						
Reading.....	0.0006	0.0000	0.0008	0.0002	0.0002	
Pipetting.....	.0000	.0006	.0004	.0010	.0010	
Chance.....	.0023	.0026	.0026	.0023	.0028	
Total.....	.0029	.0032	.0038	.0035	.0040	
	<u>Percent</u>	<u>Percent</u>	<u>Percent</u>	<u>Percent</u>	<u>Percent</u>	
Estimated percentage variance which could be eliminated by using only:						
1 reader.....	20.7	0	21.1	5.7	5.0	
1 pipetter.....	0	18.8	10.5	28.6	25.0	
	Market 9, 1 pipetter, 36 samples each read by--					
	7 readers	5 readers	7 readers	8 readers	8 readers	15 readers
	<u>Fat %</u>	<u>Fat %</u>	<u>Fat %</u>	<u>Fat %</u>	<u>Fat %</u>	<u>Fat %</u>
Within-sample variance due to:						
Reading.....	0.0003	0.0006	0.0007	0.0003	0.0002	0.0005
Chance.....	.0013	.0017	.0010	.0021	.0018	.0012
Total .....	.0016	.0023	.0017	.0024	.0020	.0017
	<u>Percent</u>	<u>Percent</u>	<u>Percent</u>	<u>Percent</u>	<u>Percent</u>	<u>Percent</u>
Estimated percentage variance which could be eliminated by using only:						
1 reader.....	18.8	26.1	41.2	12.5	10.0	29.4

because, with volume constant at 17.6 ml., a smaller weight of warm milk than of cold milk would be delivered into the test bottle. This would result in a smaller quantity of fat in the neck of the Babcock test bottle and a lower butterfat test reading. Lower viscosity at high temperatures might offset the change in volume, by causing less fat to adhere to the walls of the pipette. Consequently more complete delivery of the pipetted sample into the test bottle would occur than at lower temperatures. One purpose of this work was to determine whether pipetting temperatures are a source of downward bias in composite tests.

In some laboratories fresh samples are pipetted with no attempt to standardize the pipetting temperature. Differences in pipetting temperatures that could occur in the absence of standardizing, theoretically, could cause day-to-day variation in a producer's tests and explain part of any differences between tests on individual samples, and between plant tests and check tests. This suggested a second purpose of the work on pipetting temperatures, to determine whether differences in fresh tests occurring under the usual range of temperatures found in plants and laboratories would be significant.

Table 5.--Pipetting temperatures for milk samples specified in testing procedures required in various States, 1953 1/

Pipetting temperatures (°F.)	Number of States	Pipetting temperatures (°F.)	Number of States
50° - 70° .....	2	85° - 100° .....	1
50° - 100° .....	1	90° .....	1
55° - 65° .....	1	95° - 100° .....	2
55° - 70° .....	2	100° .....	3
60° - 68° .....	1	Cool to 70° .....	1
60° - 70° .....	7	About 70° .....	4
60° - 100° .....	1	Not over 110° .....	1
65° - 75° .....	1	Warm .....	1
68° .....	3	Not specified .....	14
70° - 95° .....	1		

1/ (2, p. 13.) Apparently these pipetting temperatures apply to both fresh and composite samples.

Work by the U. S. Bureau of Standards shows that if milk with 4.0 percent butterfat is assigned a volume of 1.0 at 68° F., the volume would be 1.0020 at 80° F., 1.0040 at 90° F., and 1.0065 at 100° F. (21). This difference in volume results in an amount of fat delivered into the testing bottle at 100° F. equal to 99.35 percent of the amount delivered at 68° F., assuming that surface tension would be equal at both temperatures. Since the surface tension of milk is lessened at higher temperatures, it is probable that the weight of milk delivered at 68° and 100° would be closer than the relation indicated above (20).

The effect of pipetting temperature on test results has been measured in several experiments. Wilster and Robichaux found that there was no difference between the averages of tests on 12 samples pipetted at 68° F. and at 80° F. A pipetting temperature of 100° F. gave an average fat reading which was 0.05 percent lower than that for 68° F., and 120° F. gave an average fat reading which was 0.08 percent lower than that for 55° F. (21). Dahlberg found the average fat reading of tests on 6 samples pipetted at 120° F. was 0.01 percent higher than the average of tests on the same samples pipetted at 70° F. (3). Bailey found that the average weight of milk delivered at 70° F. was 17.937 grams and at 115° F. was 0.123 gram less. He stated that "on the average reading of 4.51 percent this would amount to 0.046 percent" (2). A bias of only 0.02 percent butterfat, however, could cause a difference of 1 point in 20 percent of tests. Therefore, it seemed necessary to carry out additional work on the effect of pipetting temperatures on test results.

In order to obtain clear evidence on the effect of pipetting temperature on test results, 185 fresh and composite samples from five markets were each pipetted at five different temperatures, 60°, 70°, 80°, 90°, and 100° F., and each subsample was tested. The results of these tests for all markets were pooled and analyzed to determine if temperature of pipetting has an effect on the fat test.

Although statistically significant differences frequently occurred among the tests on the samples pipetted at each of the five pipetting temperatures and between the tests on the sample pipetted at 70° and 100° F., there was no consistent tendency for the



higher pipetting temperature to give lower or higher tests (tables 6 and 7.)

Both of the experiments made in Market 9 showed a consistent and highly significant inverse relation between pipetting temperatures and test results (table 7). No other market had results that showed a consistent and firm relationship. This suggests that a relationship between pipetting temperatures and test results might be difficult to establish under industry conditions.

### Storing and Reheating Composite Samples

Thirty-six States require that dairy plants retain composite samples after the end of the compositing period for times ranging from 1 to 12 days (9). The holding time affords regulatory agencies an opportunity to verify the accuracy of dealers' tests on such samples. Some groups have objected to the regulations on the basis that butterfat samples deteriorate with time so that results of check-tests made after the compositing period are inaccurate.

Demonstration that either storage time or heating and cooling significantly affect the level of test results would furnish an objective basis for reconsidering regulations required for storage of composite samples after the end of the compositing period, or for establishing tolerances between results of check-tests and the initial composite tests.

Studies comparing results of tests on fresh samples with results of tests on 7-day, 10-day, and 15-day composites showed that test results for composite samples tended to be lower than the average of results for fresh samples for the same day, and that the spread between fresh and composite tests tended to increase with the number of days in the compositing period (10, 17). This suggests that storage time may affect results of the Babcock tests.

After the original test is made on a composite sample within 24 hours after the end of the compositing period, samples to be held for retesting are immediately cooled and stored in a refrigerator. Before a second test can be made, the sample must be taken out of storage and reheated to the appropriate pipetting temperature (95°-100° F. for all tests in these experiments).

In order to find out the effects of reheating a composite sample for a retest after the compositing period, eight controlled experiments were set up in four markets. In each of these experiments milk samples were obtained for 24 producers, and composite samples were prepared for each producer. After testing at the end of the compositing period (treatment A), the remainder of each original composite sample was divided into 3 parts which were cooled and stored for retesting as follows: reheated once for testing after 1 day (treatment A<sub>1</sub>); 2 days (treatment B<sub>1</sub>); and 5 days (treatment C<sub>1</sub>). The first subsample was cooled and reheated a second time for testing after 2 days (treatment B<sub>2</sub>), and cooled and reheated a third time for testing after 5 days (treatment C<sub>3</sub>).

Tests of composite samples made, 1, 2, and 5 days after the end of the compositing period were appreciably lower in butterfat than tests on the same samples at the end of the compositing period (tables 8 and 9). About 95 percent of the tests made 1 day after the end of the compositing period (treatment A<sub>1</sub>) were equal to or less than the tests at the end of the compositing period, as were 93 percent of the tests made 2 days after the end of the compositing period (treatments B<sub>1</sub> and B<sub>2</sub>), and 99 percent of the tests made 5 days later (treatments C<sub>1</sub> and C<sub>3</sub>).

Table 6. Average fat tests and significance of their differences for fresh and composite samples of milk pipetted at five temperatures and differences between tests at 70° and 100°F with an indication of significance

Type of sample and location of testing	Samples pipetted at each temperature	Pipetting temperature					Level of significance of differences among the 5 averages 1/	Average differences between tests on samples pipetted at 70° and 100° F (70° minus 100°F) 2/
		60°	70°	80°	90°	100°		
	Number	Fat %	Fat %	Fat %	Fat %	Fat %	Percent	
Fresh:								
Market 1	20	3.84	3.83	3.83	3.82	3.82	5	+0.0075
Market 2	20	4.48	4.48	4.55	4.53	4.52	1	+0.0175 *
Market 8	20	3.87	3.90	3.90	3.90	3.89	1	-0.0350
Total	60	4.06	4.07	4.09	4.08	4.08	n.s.	-0.0033
7-day composites:								
Market 1	20	4.19	4.16	4.20	4.16	4.22	1	-0.0575 **
Market 8	20	3.86	3.87	3.87	3.88	3.87	n.s.	+0.0000
Total	40	4.02	4.01	4.04	4.02	4.04	n.s.	-0.0288 **
10-day composites:								
Market 2	20	4.48	4.48	4.46	4.46	4.46	n.s.	+0.0150
Market 9	20	3.88	3.87	3.86	3.86	3.85	1	+0.0225 **
Total	40	4.18	4.18	4.16	4.16	4.16	n.s.	+0.0188 **
15-day composites:								
Market 7	25	3.83	3.80	3.79	3.80	3.80	n.s.	-0.0040
Market 9	20	3.88	3.88	3.88	3.88	3.86	5	+0.0200 **
Total	45	3.85	3.84	3.83	3.83	3.83	n.s.	+0.0067
All samples	185	4.03	4.02	4.03	4.02	4.03	n.s.	-0.0016

\* Indicates difference statistically significant at 5-percent level.

\*\* Indicates difference statistically significant at 1-percent level.

$\frac{1}{5}$ / As determined by analysis of variance, each percentage indicates the number of times in 100 such experiments that one could expect by chance alone to find differences as large or larger if in fact there is no difference associated with pipetting temperature. The notation "n.s." (not significant) is used where the expectation is greater than 5 in 100.

$\frac{2}{5}$ / The significance of the differences for each group was determined by Duncan's Multiple Range Test and for each type of sample by the t-test.

Table 7.--Rank, by type of sample, of average butterfat test on samples of milk pipetted at different temperatures and significance and direction of straight-line relationships between average test and pipetting temperature 1/

Item	Fresh samples		7-day composite samples		10-day composite samples		15-day composite samples		All types of			
	Market		Market		Market		Market		Market			
	All		All		All		All		All			
	1	2	1	2	1	2	1	2	1	2		
Samples.....number:	20	20	20	20	20	20	40	40	20	25	45	185
Pipetting temperature												
60°F.....rank:	1	5	5	3	5	3.5	1	1.5	1	2	1.0	2.0
70°F.....rank:	3	4	1	5	3	5.0	2	1.5	3	1	2.0	4.5
80°F.....rank:	2	1	2	2	2	1.5	5	4.0	5	3	4.0	2.0
90°F.....rank:	5	2	3	4	1	3.5	4	4.0	4	4	4.0	4.5
100°F.....rank:	4	3	4	1	4	1.5	3	4.0	2	5	4.0	2.0
Direction of trend:												
line 2/ .....	-	+	+	+	+	+	-	-	-	-	-	-
Level of significance of straight-line relation-ship 3/...percent:	1	1	n.s.	n.s.	5	n.s.	n.s.	n.s.	1	n.s.	1	n.s.

1/ A rank of 1 indicates the highest average butterfat test and a rank of 5 indicates the lowest average test for the 5 temperature groups.  
2/ Minus sign indicates a drop in average butterfat test as temperature increases. Plus sign indicates an increase in average test as temperature increases.  
3/ Each percentage indicates the average number of times in 100 such experiments one could expect to find by chance alone a straight-line relationship such as the one obtained if in fact there is no linear relationship between test and pipetting temperature. The notation n.s. (not significant) is used where the expectation is greater than 5 in 100.



Table 8.--Average difference in butterfat test of composited milk samples, by days held and times reheated

Treatment	Days held	Average difference from original composite test (treatment A) <u>1/</u>		
		All composites	7-day composites	10-day composites
		<u>2/</u>	<u>2/</u>	<u>2/</u>
		Number	Butterfat percent	Butterfat percent
Reheated once:				
A <sub>1</sub> .....	1		-0.03 r	-0.03 r
B <sub>1</sub> .....	2		- .04 rs	- .05 s
C <sub>1</sub> .....	5		- .07 t	- .07 t
Reheated twice:				
B <sub>2</sub> .....	2		- .05 st	- .07 t
Reheated three times:				
C <sub>3</sub> .....	5		- .13 u	- .13 u

1/ Letters "r" - "u" indicate statistical significance. Average difference followed by letter "r" is significantly different from those differences in the same column not having "r"; those followed by "s" are significantly different from those not having "s", etc.

2/ Each average difference has been shown by a t-test to be very highly significant (except for the 2 in 7-day column marked with asterisks to indicate no significance) that is, on the average, the test on each sample tested after the end of the compositing period was lower than the original composite test by an amount which could be expected to occur in not over 1 percent of the trials due to chance alone.

The number of times a sample was reheated was a more important factor than the length of storage. Of the three sets of tests made with only one reheating, only those held 5 days (treatment C<sub>1</sub>) were significantly different from the other two sets. On the average, treatment C<sub>1</sub> resulted in butterfat tests about 0.07 percent lower than the original composite (treatment A) (table 8).

The samples held for 1 day (treatment A<sub>1</sub>) and those held for 2 days (treatment B<sub>1</sub>) averaged lower than the original composite by 0.03 and 0.04 percent. They were close enough to each other, however, to represent differences which had a high probability of occurrence due to chance, and the effects of the two treatments could not be considered to be different.

The tests made on samples held for 2 days and reheated twice (treatment B<sub>2</sub>) averaged 0.05 lower than the original composite test. This is not significantly different from the average for treatment B<sub>1</sub>, 0.04, held the same length of time but reheated only once, or from the average of treatment C<sub>1</sub>, 0.07, held 5 days but reheated only once.

The samples differing the most from the original composite tests were those held 5 days and reheated three times, (treatment C<sub>3</sub>). Their average difference of 0.13 was significantly lower than the differences for any of the other four types of treatments (table 8).

The downward effect on test results of reheating suggests that allowing composite samples to stand at room temperature during any part of the compositing period is a

Table 9.--Average difference between butterfat test after compositing period and the regular composite control test (treatment A) by markets 1/

Item	7-day composites					10-day composites				
	7- and 10-day		Market		All samples	Market		Market		Market
	composites	combined	1	5		1	5	1	5	
Treatment A <sub>1</sub> :										
Average difference from A.....Pct. fat:	-027	-024	-023	-025	-028	-058	-012	-033	-038	0.000
Total 2/.....Samples:	192	48	24	24	144	24	24	24	24	24
Higher fat test 3/.....Samples:	9**	5	1	4	4**	0*	0	0	0	4
Same fat test.....Samples:	119	23	12	11	96	11	21	16	15	24
Lower fat test.....Samples:	64	20	11	9	44	13	3	8	9	11
Treatment B <sub>1</sub> :										
Average difference from A.....Pct. fat:	-039	-010	-054	+033	-049	-083	-042	-071	-025	-031
Total 2/.....Samples:	192	48	24	24	144	24	24	24	24	24
Higher fat test 3/.....Samples:	15**	12	1**	11	3**	0**	0	0**	0	3
Same fat test.....Samples:	86	17	8	9	69	7	14	7	18	9
Lower fat test.....Samples:	91	19	15	4	72	17	10	17	6	12
Treatment C <sub>1</sub> :										
Average difference from A.....Pct. fat:	-066	-060	-033	-088	-068	-092	-058	-075	-054	-073
Total 2/.....Samples:	192	48	24	24	144	24	24	24	24	24
Higher fat test 3/.....Samples:	3**	3**	2*	1**	0**	0**	0**	0**	0*	0**
Same fat test.....Samples:	69	15	8	7	54	6	11	6	12	8
Lower fat test.....Samples:	120	30	14	16	90	18	13	18	12	16
Treatment B <sub>2</sub> :										
Average difference from A.....Pct. fat:	-054	-016	-019	-012	-067	-096	-058	-083	-075	-048
Total 2/.....Samples:	192	48	24	24	144	24	24	24	24	24
Higher fat test 3/.....Samples:	11**	10	5	5	1**	0**	0*	0**	0**	1*
Same fat test.....Samples:	74	20	8	12	54	4	11	6	10	9
Lower fat test.....Samples:	107	18	11	7	89	20	13	18	14	14
Treatment C <sub>3</sub> :										
Average difference from A.....Pct. fat:	-134	-151	-229	-062	-128	-167	-138	-096	-133	-110
Total 2/.....Samples:	189	45	24	21	144	24	24	24	24	24
Higher fat test 3/.....Samples:	1**	1**	0**	1*	0**	0**	0**	0**	0**	0**
Same fat test.....Samples:	27	10	2	8	17	1	5	5	1	2
Lower fat test.....Samples:	161	34	22	12	127	23	19	19	23	22

1/ The regular composite sample, treatment A, was tested within 24 hours after the end of the compositing period. The composite sample was then divided into sub-samples which were cooled immediately and stored in the refrigerator. After 24 hours, one sub-sample was heated and sample for treatment A<sub>1</sub> was pipetted and tested. After 48 hours, the same sub-sample was reheated and sample for treatment B<sub>2</sub> was pipetted and tested, and after 5 days the sub-sample was again reheated and the sample for treatment C<sub>3</sub> was pipetted and tested. The sub-sample was cooled immediately after each pipetting. The second sub-sample for treatment B<sub>1</sub> was heated, pipetted, and tested 48 hours after test of treatment A; and the third sub-sample for treatment C<sub>1</sub> was heated, pipetted and tested 5 days after test of treatment A. The pipetting temperature was 95°-100°F for all samples.

2/ Number of samples is the number of producer deliveries for which the six types of composite tests were made.

3/ The significance of the number of times the less frequent plus or minus sign occurred has been tested by the statistical sign test of Dixon and Mood. Ref: Journal of American Statistical Association, v. 41, no. 236, (Dec. 1946) pp. 557-566).

\*Indicates 5 percent level of significance.

\*\*Indicates 1 percent level of significance.



possible cause of downward bias in composites as compared with results on fresh samples.

The averages in table 8 afford check-testing agencies some measure of the tolerances appropriate when check-testing is delayed after the compositing period has ended.

## THE BLENDING OF CAN MILK IN PLANT WEIGH TANKS

In most plants receiving milk in cans, samples for Babcock and other tests are taken from the weight tank as the cans from each producer's delivery are dumped and weighed. Should the milk be inadequately blended in the weigh tank before the sample is taken, it may not be representative of the producer's total delivery.

Weigh tanks may vary considerably in their ability to blend milk. Samples taken at any one place in the tank may not be representative of the entire contents of the tank. In markets where plants account to producers' organizations or market administrators for milk intake and fat tests, weigh tanks usually are required to meet standards for mixing ability. Nevertheless, under normal operating conditions, differences occur among weigh tanks in their blending of milk and these can affect butterfat test results. The amount of such variability could cause significant differences in results from two samples of the same milk, where each was taken from a different place in the weigh tank (17, 1, 14, 15, 16, 19).

This consideration led to analysis of the blending ability of weigh tanks used in eight plants where butterfat tests were made for the present study. For this limited study of blending ability, one series of samples was taken from the place in the tank that the plants ordinarily used. Samples were also taken from one to four other places in the tanks. Test results for the samples from each position in the tanks were compared and were analyzed statistically to determine whether differences were greater than could be expected by chance alone.

For weigh tanks where some method of agitation was followed to improve the blending ability of the tanks before sampling, from 0 to 36 percent of the samples from one position differed in test by more than 1 point from samples from another position (table 10). These percentages varied sharply among tanks with the same method of agitation, (Plants 1 and 4; 12 and 14). One weigh tank showed a relatively high proportion of tests (36 percent) differing by more than 1 point, and the largest average difference between positions, 0.0764. Statistically, this average difference was not significantly different from a zero difference, and did not represent a "bias" between the two positions in the tank. The average difference, though large, could not be considered significant because it is not larger than one would expect on the basis of the variation in the size of the individual differences. 6/

In 5 plants weigh tanks were sampled without previously agitating the milk. For these tanks, from 10 percent to 50 percent of the tests on samples from one position differed by more than 1 point from those on samples from a second position (table 10).

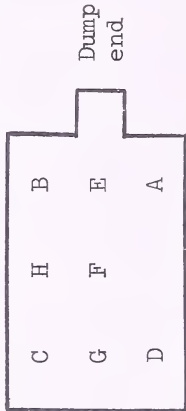
---

6/ In this study the designation, "significant difference", means that differences as large or larger than the one occurring would be expected by chance alone in not more than 5 percent of repeated trials. "Highly significant difference" means that differences as large or larger than the one occurring would be expected by chance alone in not more than 1 percent of repeated trials.

Table 10.--Differences in butterfat test results between milk samples taken from two positions in plant weigh tanks

Agitation method (limited studies only) and plant number	Positions compared 1/ differences	Number of paired differences	Percent of total number of differences			Average size of difference 2/ of difference 2/
			Less than : 1/2 point		Over 1 point	
			Percent	Percent	Percent	
Butterfat percentage						
<hr/>						
No agitation:						
Plant 6	C-D	30	27	43	30	-.0383
Plant 7	C-D	30	23	67	10	.0267
Plant 8	C-D	30	40	40	20	-.0517**
Plant 14	E-G	22	14	36	50	-.0386*
Plant 15 3/	F-E	20	55	35	10	.0200
Plant 15 3/	G-E	20	25	50	25	.0050
Blending chute:						
Plant 1	H-G	36	17	47	36	.0764
Plant 4	C-D	30	60	37	3	.0100
Plant 4	F-D	30	63	37	0	-.0027
Hand:						
Plant 12	A-D	52	79	19	2	-.0125
Plant 14	E-G	12	34	33	33	.0333

1/ The sampling positions are approximately identified in the following general diagram:



2/ Asterisks indicate level of significance of "bias" as determined by a t-test. In repeated trials, equal or greater average differences could be expected to occur by chance in no more than:

1 percent of the trials (\*\*) or  
5 percent of the trials (\*).

3/ A "milk thief" was used to take samples.

In one plant where the experiment was done first without agitation, a second trial was made, taking samples after hand agitation (Plant 14). After agitation, 33 percent of the differences were greater than 1 point and the average difference of 0.0333 was not significant; without agitation, 50 percent of the differences were greater than 1 point and the average difference of -0.0386 was significantly different from a zero average difference.

When tests on samples from two positions are in fact equivalent except for random, or chance, variations in sampling and testing, the average difference in the paired tests can be expected to fluctuate around zero. Average differences obtained were tested by a t-test to determine the probability of occurrence due to chance, on the basis of the variation of the individual differences, of an average as large as or larger (and, therefore, as different from zero) than the one obtained. When the probability is 5 percent or less, the average difference is considered to be significantly different from zero and to represent a "bias" between the two positions being compared. The two positions being compared differed significantly in two of the four experiments where samples were taken without agitating the milk, and without the use of the "milk thief," and in none of the five trials where samples were taken following some form of agitation.

The percentage of samples from two positions disagreeing by more than one point is probably a better measure of the mixing ability of weigh tanks than the significance of the average difference. Plus and minus differences between samples taken from different positions may balance each other so that the average difference does not represent a statistically significant "bias," but either plus or minus differences of over one point would reflect incomplete mixing. When milk has been thoroughly blended, there is no reason to expect tests on samples from two positions to vary more than do two tests by one tester on a single sample. Two tests by the same tester on one sample of milk can be expected to differ by more than one point in less than 2 percent of the comparisons. (See pages 9 and 10.)

## PROBLEMS RELATED TO PICK UP OF MILK IN BULK TANKS

The widespread adoption of bulk handling of milk has brought with it a need to develop a system for sampling and testing milk that protects both plant and producer, at an acceptable cost, against added variability in butterfat tests. This concern refers to variations other than the inherent day-to-day change in butterfat content. Under the bulk tank system of hauling, farm milk is commingled at the farm instead of the plant. This makes it necessary to take milk samples for butterfat testing at each farm, before the milk is pumped from the farm tank into the tank truck. Bulk tank milk usually is sampled in one of two ways: (a) A sampler rides the tank truck and takes a sample at each farm, or (b) the driver of the truck takes the sample.

The first way of sampling is expensive and usually impractical because one sampler can visit relatively few farms daily. It is doubtful if many plants would consider taking samples, especially daily samples for compositing, in this high-cost fashion. On the other hand, it may be necessary and practical to obtain random fresh samples for check-testing in this way.

The least costly way to sample milk in farm bulk tanks is to have the hauler take the sample. However, the representativeness of the sample, if taken in this way, may be questioned on the basis that the hauler is not necessarily a skilled sampler, that he may be careless in taking samples, or he even may be suspected of deliberately taking unrepresentative samples. In some markets, haulers are required to pass State tests and be licensed as milk samplers in order to ensure that they understand and can follow



acceptable methodology in taking samples. However, licensing a sampler does not ensure that he will sample correctly so a need persists to afford producers and plants assurance against erroneous sampling.

Samples also may become unrepresentative through damage during the transportation from farm to plant. Samples may churn or freeze, so that composite samples built at the plant are unrepresentative. Most bulk tank trucks have insulated sample compartments. Sample bottles also are iced to minimize the possibility of damage to samples during the warm months. However, even with these safeguards there is chance of damage to the samples.

Plants and check-testing agencies have two concerns regarding sampling of bulk milk at the farm: (1) They need a way to check a hauler's sampling without incurring the expense of an official sampler on each bulk tank truck; (2) They need to determine which possible causes of unrepresentative samples result in differences so small that they can be ignored.

The following studies were made to assemble information about these problems and to evaluate ways which have been suggested or used to meet them.

### Sampling from Bulk Tank Trucks

One way to minimize the cost of checking the hauler's sampling and the plant's testing of bulk tank milk is to use samples taken by the tank truck drivers and samples from the loaded tank truck at the plant. Samples taken from the loaded tanker, if representative of the milk in the tanker, presumably contain the percentage of fat equal to the average of the fat tests of all producers whose milk is in the tank, weighted by the pounds of milk each one delivered. Therefore, if the fat test of the tanker milk equals the weighted average of all producers' tests determined from the driver's samples, within appropriate tolerances, the check-testing agency would assume that the tests of the individual producers' milk were accurate and the drivers' samples were representative. This assumption does not rule out compensating errors, since a low test for one producer might be balanced by a high test for another. However, statistically determined tolerances afford a testing laboratory a most helpful guide for detecting improper sampling of individual producers' milk. These tolerances would represent expectations based on results obtained by unbiased testers working under normal commercial laboratory conditions. The method appears to afford a possibility of maintaining, at low cost, a constant check on the accuracy of sampling and testing when it is used in conjunction with periodic check-testing of individual producer's samples. In 1957 this practice was followed in eight Federal Order markets.

This method of verification presupposes knowledge of the differences to be expected in test results between samples of milk taken properly from bulk tank trucks and the weighted averages of tests on proper samples of the individual producers' milk.

Because the representativeness of the sample obtained from the loaded tank truck may vary from one sampling method to another, the amount and dispersion of differences can vary. The amount of agitation given to the milk in the tank truck would be expected to affect the representativeness of the sample.

A study was made to measure the amount of differences between the weighted averages of tests on samples from individual producers' milk and tests on samples of their commingled milk from bulk tank trucks. These experiments, made in three

different markets, compared the representativeness of results from sampling in several ways both with and without agitating the tank load before sampling.

The agreement between the test of samples from a bulk tank truck and the weighted average of tests of producers' samples was close when the tank load of milk had been agitated (experiments 1 and 4, manhole sampling, and experiment 4, inline sampling using automatic positive periodic sampler and samples from the plant's holding tank, table 11).

Since a tank load of milk is agitated at least partially by pumping producers' milk into the tank truck, the time between pumping the last milk into the tank truck and the sampling tended to affect the agreement between the test of the commingled milk in the tank truck and the weighted average of producers' tests when no further agitation was given the milk in the tanker before sampling. This can best be seen by comparing the differences for experiment 5 with those for the other experiments when there was no agitation. However, among individual tank loads, the relation of time to agreement differed widely. Undoubtedly factors such as size of fat globule and viscosity of milk in individual loads, which affect creaming time, and condition of roads, which could affect "surge" and therefore agitation, and the volume of the last pick-up in relation to the volume of the milk already in the tank, modify this relationship.

The size and dispersion of differences between tests on milk from the tank trucks and the weighted average of producers' tests varied considerably from one method of sampling to another. In experiment 2, with no agitation, the manhole samples tended to test about a halfpoint high because the milk had started to cream. A high proportion of samples from the valve at the bottom of the tank tested very low in this experiment, because the milk had started creaming.

The effect of creaming is shown more definitely by the valve samples in experiment 5. Most samples taken at the beginning and during the middle of the unloading tested low; while a large proportion of the samples taken at the end of the unloading, and therefore from the top of the tank, tested very high.

In experiment 4, all sampling methods agreed closely (table 11). This may be explained by the short time lapse between pumping the milk last picked up into the tank truck and the sampling time.

The average size and direction of differences for each sampling method, with notation as to agitation, also are shown in table 11, and the methods which resulted in statistically significant biases from the weighted averages of producers' tests are identified. The fairly wide average difference, -0.0215 percent butterfat, for manhole sampling without agitation in experiment 1, was not statistically significant because the differences for the five individual tankloads varied so much that an average difference this large or larger had a probability of about 25 percent of occurring on the basis of chance. On the other hand, in experiment 3b an average difference of about the same size, -0.0207, was statistically significant because the differences were consistently below the average of the individual producers' tests. This consistent difference in one direction would have less than a 1 percent probability of occurring on the basis of chance alone.

The results of these experiments indicate that reliably representative samples can be obtained from loaded tank trucks only if the milk has been agitated before sampling, regardless of the method of sampling. In some circumstances the amount of agitation afforded by pumping producers' milk into the tank may be sufficient.



Table 11.--Differences in butterfat tests between milk samples taken from bulk tank trucks and the weighted averages of tests on individual producers' milk

Experiment number and average time, last pick-up to sampling	Number: of tank loads	Number: of tank loads per sample	Sampling method	Agitation difference: butterfat percentage	Average : average of producer tests, by butterfat percentage	: Percent of bulk tank samples differing from weighted
No. 1, 19 minutes.....	5	2	Manhole	None	-0.0215	0 0 0 20 80 0 0 0 0
		2	Manhole	Yes	0.0035	0 0 0 0 100 0 0 0 0
No. 2, 27 minutes.....	15	4	Manhole	None	0.0856**	0 0 0 0 27 46 20 0 7
		4	Drip (cracked valve)	None	0.0072	0 0 0 6 12 67 15 0 0
		2	Valve at bottom of tanker, 8-oz. sample, beginning of unloading	None		
No. 3(a), 24 minutes....	45	4	Drip (cracked valve)	None	-0.3356**	47 0 13 0 33 7 0 0 0
No. 3(b), No record....	29	4	Drip (cracked valve)	None	-0.0061	0 0 9 11 62 18 0 0 0
No. 4, 15 minutes.....	31	2/1	Manhole	None	-0.0207**	0 0 3 10 84 3 0 0 0
		2/1	Manhole	None	-0.0008	0 0 6 0 94 0 0 0 0
		2/1	Manhole	Yes	-0.0063*	0 0 0 3 97 0 0 0 0
		2/1	Inline (automatic positive periodic sampler)	Yes	-0.0028	0 0 0 0 100 0 0 0 0
		2/1	Holding tank in plant	Yes	3/-0.0081*	0 0 0 0 100 0 0 0 0
No. 5, 122 minutes.....	4/59	4	Inline (automatic positive periodic sampler)	None	-0.0679**	5 10 20 31 32 0 0 0 2
		4	Valve (beginning of unloading)	None	-0.4584**	28 5 10 18 35 2 0 0 2
	4/59	4	Valve (middle of unloading)	None	-0.1899**	29 7 17 25 22 0 0 0 0
	60	4	Valve (end of unloading)	None	1.0582**	0 7 17 18 12 2 0 3 41

1/ Test on tank sample minus weighted average of producers' tests. Asterisks indicate the level of significance of the average differences. In repeated trials, equal or greater average differences could be expected to occur by chance in no more than; one percent of the trials (\*\*) or five percent of the trials (\*). 2/ One test was made of each sample and each test was read 8 times. 3/ In this experiment small differences, .04 or less, were significant at the 5 percent level because they tended to be in one direction. 4/ One sample for one tank load for this method was lost.

In general the use of loaded tanker testing as a check-test appears to have certain useful applications. (1) Study of table 11 shows that when a tank load of milk has been agitated, its test can be expected to agree within 1 point of the weighted average of tests of individual producers' milk in the tanker, in over 95 percent of trials. Therefore, a difference as large as 0.2 would be expected in less than 5 percent of tank loads. (2) For any bulk-tank route it is possible to make a frequency distribution of differences between the loaded tanker test and the average of producers' tests, over a number of comparisons. A frequency distribution of this kind could be compared with results shown in table 11 to determine whether the results were in reasonable agreement. For example, on the basis of tests on samples taken from tanker manholes, after agitation, for 36 tank loads in experiments 1 and 4, the tank sample can be expected to test within 0.04 percent of the weighted average of producers' tests in about 95 percent of the trials. The probability that a test will differ from the weighted average by 0.04 to 0.09 is between 4 and 5 percent. A difference of one point or more would be expected in less than 1 percent of the comparisons.

### Adding Preservatives to the Farm Sample

At the time this study was undertaken, it was the practice in one of the markets to add 4 drops of a 36-percent mercuric chloride solution to Babcock test bottles as a preservative before pipetting the duplicate samples, in case a retest became necessary. This eliminated the development of a sour smell, (that is, prevented bacteriological deterioration of the sample) and made the test bottles easier to clean when the duplicate was held several days.

The 1955 edition of the Official Methods of Analysis of the Association of Official Agricultural Chemists recommends that a "Tablet containing  $\text{HgCl}_2$  (mercuric chloride),  $\text{K}_2\text{Cr}_2\text{O}_7$  (potassium dichromate), or other suitable preservative, weighing not more than 0.5 gram for 8 fluid ounces of milk, or 36 percent solution of  $\text{HCHO}$  (formaldehyde), 0.1 milliliter (2 drops) per fluid ounce, may be used..." (12). An ounce of milk is about 30 ml., and the Babcock test requires 17.6 ml. of milk. Therefore the use of 4 drops of solution with each pipetted sample gives a much larger amount of solution per unit of milk than is recommended for composites.

Indications from previous research are that composite samples tend to test lower than fresh samples. It was not known whether the use of excess amounts of  $\text{K}_2\text{Cr}_2\text{O}_7$  would affect test results. For this reason, the present study was undertaken to determine if pipetted samples to which four drops of a preservative had been added would test significantly different from samples to which no preservative had been added. This analysis was made on duplicate fresh samples taken from 50 farm bulk tanks.

Tests on samples to which preservatives had been added averaged 0.009 percent butterfat above tests on corresponding samples without preservatives. On the basis of the variance of the individual sample differences, an average as large as or larger than 0.009 had a probability of occurrence of from 20 to 30 percent and would not indicate a significant difference between the samples with preservative and those with no preservative. Tests on 86 percent of the paired samples agreed, and the remaining 14 percent differed by 1 point. This is very close agreement with the average differences on identical, or split, samples of milk tested by pairs of technicians who were accustomed to working together: 74 percent in agreement and 26 percent differing by 1 point or more.

## Transporting Samples From the Farm to the Laboratory

When milk samples are taken from farm bulk tanks and transported to the laboratory for testing, the motion of the truck may churn the samples and cause loss of butterfat in the fat test. Larger particles of butter in churned samples cannot be drawn into the pipette; hence, the milk delivered into the test bottle may be lower in fat than the milk from which the sample was drawn.

Research by Ragsdale and others showed only "slight differences" in tests on 17 composite samples built and held at a laboratory and 17 duplicate composite samples which were built at the farm but transported every other day to and from the farm in the refrigerated sample compartment of a tank truck (18). That research, however, does not throw light on the effect on test results of transporting fresh samples because the composites were transported rather than the daily samples used to build the composites.

This study was initiated to determine whether the transportation of fresh samples, from the farm to the laboratory affected the level of test or the variability of producers' tests, either in testing fresh samples or composite samples. In one market, part of each of 315 samples taken for fresh tests was heated to 68° F. and pipetted into test bottles at the farm. This pipetted part of the sample and the remainder of the sample were taken to the market administrator's laboratory in the sample compartment of the tank truck. There, milk from the remainder of each sample was pipetted into test bottles, and tests were made on the duplicate samples pipetted at the farm and laboratory. In a second experiment in the same market, 158 fresh samples were taken from farm bulk tanks and pipetted at the farm by the market administrator's technicians. A second set of samples was collected at the farms by the driver of the tank truck, brought to the plant, and pipetted by the market administrator's technicians.

In addition to the 2 experiments comparing fresh tests, 10 experiments were made with 164 pairs of composite tests. In this part of the work duplicate composite samples were built for each producer whose milk was tested. One of these was built at the farm, as the sample was taken, and kept at the farm. The rest of the daily sample was taken to the laboratory and added to the second composites which were held at the laboratory. The composite held at the farm was pipetted there before being taken to the laboratory, where all of the samples were tested.

In 7 of the 12 experiments, test results on the samples pipetted at the farm differed significantly from results on samples pipetted at the laboratory (table 12). Damage to samples from transportation would be expected to cause consistently lower tests on the samples pipetted at the laboratories. Such consistency did not occur in the 7 experiments where differences were significant. In 4 of the 7 experiments, farm-pipetted samples tested significantly higher than the samples pipetted at the laboratory; in the other 3, farm-pipetted samples were below the laboratory samples. The combined data for each of the 3 kinds of samples used in the experiments--fresh samples, 7-day composites, and 15-day composites--showed no significant difference in test results between farm pipetting and laboratory pipetting (table 12). These experiments were performed in February, March, April, and November. No relationship existed between the month and size of difference. These experiments do not indicate any damage to samples in connection with transportation.

Samples in these experiments were carefully handled, since drivers were aware that duplicate tests were being made. It is entirely possible that individual drivers, particularly in hot weather, may damage samples by improper handling on the truck. This work has not measured the effect on tests of improper handling. To do so would



Table 12.--Number of samples pipetted at the farm that tested higher and lower in butterfat than samples pipetted at the laboratory

Type of sample	Farm samples			Farm samples			Average difference, farm sample minus laboratory sample <sup>1/</sup>
	lower in butterfat than laboratory samples by--			higher in butterfat than laboratory samples by--			
	No difference			No difference			
	2 points and more	1 point		1 point	2 points and more		
	Number	Number	Number	Number	Number	Fat percent	
Fresh samples:							
315 pairs.....	0	6	294	15	0	+0.008**	
158 pairs.....	5	36	77	35	5	- .001	
Total, 473 pairs...	5	42	371	50	5	+ .005	
	Percent	Percent	Percent	Percent	Percent		
	1	9	78	11	1	--	
Composite samples:							
7-day:	Number	Number	Number	Number	Number		
9 pairs.....	0	0	3	4	2	+ .089**	
9 pairs.....	0	3	5	1	0	- .033*	
9 pairs.....	0	0	4	5	0	+ .056**	
9 pairs.....	0	3	4	2	0	- .011	
Total, 36 pairs...	0	6	16	12	2	+ .025	
	Percent	Percent	Percent	Percent	Percent		
	0	17	44	33	6	--	
15-day:	Number	Number	Number	Number	Number		
24 pairs.....	0	0	19	5	0	+ .029**	
24 pairs.....	1	6	15	2	0	- .019*	
20 pairs.....	0	3	14	3	0	.000	
20 pairs.....	0	6	9	5	0	- .002	
20 pairs.....	0	3	15	1	1	- .020	
20 pairs.....	2	7	8	3	0	- .038*	
Total, 128 pairs...	3	25	80	19	1	- .007	
	Percent	Percent	Percent	Percent	Percent		
	2	20	62	15	1	--	
All composites:	Number	Number	Number	Number	Number		
Total, 164 pairs...	3	31	96	31	3	--	
	Percent	Percent	Percent	Percent	Percent		
	2	19	58	19	2	--	

<sup>1/</sup> Asterisks indicate the level of significance of the average differences (test on farm sample minus test on laboratory sample). In repeated trials, equal or greater average differences could be expected to occur by chance in no more than: one percent of the trials (\*\*) or five percent of the trials (\*).



require controlled experiments in which duplicate tests were made, one on properly handled samples, the other on samples which had been deliberately mishandled in specific ways.

#### Including Defective Daily Portions in Building Composite Samples

One of the problems in building composite samples is whether portions should be added from daily samples which have been churned or from milk which is partly frozen. For purposes of this study three plants made a record of defective daily samples and included portions from them in some of their composite samples. For over 90 percent of the composite samples none of the daily samples had been defective. For each producer, tests on composites containing 1 or more defective daily portions were compared with averages of fresh tests for those days of the same period on which the samples were not defective. Tests on composites with no defective portions were also compared with averages of fresh tests for the period (table 13).

The distribution of the differences for both series of 10-day and of 15-day composites are shown in table 14. For each type of composite, 10-day and 15-day, the two distributions of comparisons were shown, by chi-square tests, to vary significantly at the 1-percent level.

Both 10-day and 15-day composites with defective portions had a lower proportion of comparisons agreeing within the limits  $-0.09$  to  $+0.09$  percent butterfat than the composites with no defective portions (table 14). The average differences for the two series of 10-day composites were not significantly different from each other, but for the 15-day composites they were significantly different at the 5 percent level. 7/ For both 10-day and 15-day composites, the average difference between composite and fresh samples was greater but not in the same direction (plus for 10-day and minus for 15-day composites) for composites containing some defective samples than for samples with no defective portions (table 14).

The distributions of differences for the composite tests with defective portions were influenced by two factors: (1) Varying numbers of defective daily samples during compositing periods, and (2) smaller numbers of fresh tests in the averages used in comparisons with composites which included some defective portions.

Of the 431 10-day composite samples which included defective portions, 83 percent had one portion defective, 12 percent had two, and 5 percent had three. The average number of defective portions per 10-day composite sample, and consequently the average number of fresh tests omitted from the comparable average of fresh tests for the average of 5.56 bulk tank deliveries during a 10-day period, was 1.22. Of the 109 15-day composite samples which included some defective portions, 79 percent had one portion defective, 12 percent had two, and 9 percent had three. The average number of defective portions per 15-day composite sample, representing also the average number of fresh tests omitted from the comparable average of fresh tests for the average of 8.45 bulk tank deliveries during a 15-day period, was 1.30.

---

7/ A more rigorous test of the effect of defective samples would require two series of samples, one including defective samples, the other including normal samples for all days of the testing period. Data of this kind would be difficult to obtain. It might be feasible to obtain defective and normal samples for the same lots of milk under laboratory conditions, where conditions could be controlled to induce churning or freezing after the normal sample has been drawn.

Table 13.--Average difference in butterfat between composite milk samples with and without defective portions and the sample average of fresh tests for the period 1/

Plant and compositing period	Composites including some defective portions											
	Composites with :			:			:			:		
	no defective portions :			All :			Churned :			Frozen :		
	:Av. diff., : composite : Number : minus : fresh av. :	:Av. diff., : composite : Number : minus : fresh av. :	:Av. diff., : composite : Number : minus : fresh av. :	:Av. diff., : composite : Number : minus : fresh av. :	:Av. diff., : composite : Number : minus : fresh av. :	:Av. diff., : composite : Number : minus : fresh av. :	:Av. diff., : composite : Number : minus : fresh av. :	:Av. diff., : composite : Number : minus : fresh av. :	:Av. diff., : composite : Number : minus : fresh av. :	:Av. diff., : composite : Number : minus : fresh av. :	:Av. diff., : composite : Number : minus : fresh av. :	Mixed
	Samples % fat	Samples % fat	Samples % fat	Samples % fat	Samples % fat	Samples % fat	Samples % fat	Samples % fat	Samples % fat	Samples % fat	Samples % fat	% fat
Plant 17:												
1st 10 days ...	263	0.0059	18	-.0044	13	-.0008	5	-.0140	0	---		
2nd 10 days ...	266	.0130	15	-.0033	10	.0210	4	-.0400	1	-.1000		
3rd 10 days ...	261	.0111	20	.0295	10	.0310	9	.0222	1	.0800		
Plant 18:												
1st 10 days ...	787	-.0201	70	-.0267	31	-.0342	33	-.0221	6	-.0133		
2nd 10 days ...	761	-.0255	97	-.0086	41	-.0188	53	-.0009	3	-.0033		
3rd 10 days ...	769	-.0161	89	-.0133	33	-.0252	54	-.0031	2	-.0900		
Plant 19:												
1st 10 days ...	491	.0265	35	.0488	24	.0496	11	.0427	0	---		
2nd 10 days ...	481	.0282	48	.0358	31	.0258	17	.0541	0	---		
3rd 10 days ...	491	.0320	39	.0228	18	.0183	20	.0195	1	.1700		
Plant 18:												
1st 15 days ...	443	-.0213	52	-.0517	21	-.0695	27	-.0267	4	-.1275		
2nd 15 days ...	437	-.0241	57	-.0284	23	-.0252	31	-.0232	3	-.1067		

1/ All averages of fresh tests were for nondefective samples only.

Table 14.--Percentage distribution of composite milk samples, with and without defective portions testing higher and lower in butterfat than the average of daily fresh samples, by size of differences.

Difference: composite test minus average of fresh tests <sup>1/</sup>	10-day composites		15-day composites	
	No portions defective	Including some defective portions	No portions defective	Including some defective portions
Butterfat percentage	Percent	Percent	Percent	Percent
+ .90 to +.99 .....	0.0	0.0	0.1	0.0
+ .30 to +.39 .....	<u>2/</u>	0.5	0.0	0.0
+ .20 to +.29 .....	0.1	0.2	0.0	0.0
+ .10 to +.19 .....	7.2	10.9	4.7	6.4
+ .01 to +.09 .....	39.1	34.6	28.8	18.4
0.00 .....	10.6	13.5	5.9	8.3
-.01 to -.09 .....	35.0	27.4	46.1	41.3
-.10 to -.19 .....	7.6	11.8	13.4	22.9
-.20 to -.29 .....	0.4	0.9	1.0	2.7
-.30 to -.39 .....	0.0	0.2	0.0	0.0
-.40 to -.49 .....	0.0	0.0	0.0	0.0
-.50 to -.59 .....	<u>2/</u>	0.0	0.0	0.0
Total .....	100.0	100.0	100.0	100.0
Average difference, butterfat percentage ...	.0005	.0023	-.0236	-.0395

<sup>1/</sup> All averages of fresh tests are for nondefective samples only.

<sup>2/</sup> Less than 0.05 percent.



Data for the three plants participating in this study were analyzed to determine whether, and how much, the variance of the average of fresh tests of all daily samples might differ from the variance of the averages after exclusion of defective daily samples. Table 15 indicates that for each of the periods, the reduction in the number of fresh tests per period would lead to an increase of 0.001 in the expected variance of the average of fresh tests.

The variance due to fewer fresh tests could be expected to be random, with about the same number of plus and minus variations, which would average close to zero. For this reason the average differences in butterfat shown in table 14 probably show very little effect from the smaller number of fresh tests in the averages used in comparisons with the composites built with some defective portions. The estimated increase of 0.001 in the variance of the average of fresh tests could result, for 95 comparisons out of 100, in an average being up to 0.002 higher or lower than the average computed from the normal number of tests during the period. This would be expected to influence the distribution of differences, but since 0.002 is small compared with the differences shown in table 13 this factor would not change the conclusions:

- 1) For both the 10-day and 15-day composites the distributions of differences from averages of fresh tests are significantly different for composites with no defectives and composites with some defective portions.
- 2) For 15-day composites, comparisons with averages of fresh tests show average results that differ for composites with no defectives and those with some defective portions by an amount which could be expected, due to chance alone, in not more than 5 percent of repeated trials.

Table 15.--Expected variance of averages of daily fresh tests for 10-day and 15-day compositing periods: Periods with no defectives vs. periods with some defective daily samples

Plant 1/	10-day compositing periods			15-day compositing periods		
	Average number of	Expected		Average number of	Expected	
	fresh samples and	variance		fresh samples and	variance	
	tests per period 2/	of average 3/		tests per period 2/	of average 3/	
	Number	Fat %		Number	Fat %	
17. a) No defectives .....	5.06	0.002		7.48	0.001	
b) Some samples defective .....	3.83	.003		4/	4/	
Change .....	-1.23	+ .001		---	---	
18. a) No defectives .....	5.70	.003		8.62	.001	
b) Some samples defective .....	4.43	.004		7.15	.002	
Change .....	-1.27	+ .001		-1.47	+ .001	
19. a) No defectives .....	5.09	.002		7.58	.001	
b) Some samples defective .....	4.36	.003		4/	4/	
Change .....	-0.73	+ .001		---	---	

1/ All bulk plants, averaging 5-6 pickups in a 10-day period and 7-8 pickups in a 15-day period and, therefore, 5-6 or 7-8 fresh samples per period.  
2/ Number of nondefective samples only.  
3/ Estimated standard error of the mean, based on the average number of days.  
4/ No defective portions involved in 15-day composites at this plant.

# LITERATURE CITED

- (1) Bailey, D. E.  
1919. Study of the Babcock Test for Butterfat in Milk. Jour. Dairy Sci. 2(5):331-373.
- (2) Bailey, D. H.  
1934. Methods of Sampling Milk. Pa. Agr. Expt. Sta. Bul. 310.
- (3) Dahlberg, A. O.  
1923. Comparison of the Roesse-Gottlieb and Babcock Methods of Testing. Assoc. Off. Agr. Chem. Jour. 7:159-169.
- (4) Gould, I. A., and Armstrong, T. V.  
(n.d.) Present Status of the Babcock and Other Practical Fat Tests. Ohio State Univ., Columbus, Ohio. 15 pp.
- (5) Herreid, E. O.  
1942. The Babcock Test; A Review of the Literature. Jour. Dairy Sci. 25(4):335-370.
- (6) \_\_\_\_\_  
1953. Report on Fat in Dairy Products. Variations by Different Technicians in Estimating Upper Meniscus on the Fat Column of the Babcock Test for Milk. Assoc. Off. Agr. Chem. Jour. 36(2):183-185.
- (7) \_\_\_\_\_, Burgwald, L. H., Herrington, B. L., and Jack, E. L.  
1950. Standardizing the Babcock Test for Milk by Increasing the Volume of the Sample and Eliminating the Meniscus on the Fat Column. Jour. Dairy Sci. 33(10):685-691.
- (8) \_\_\_\_\_, Burgwald, L. H., Herrington, B. L., and Jack, E. L.  
1952. Report on Fat in Dairy Products. Methods for Standardizing the Babcock Test. Assoc. Off. Agr. Chem. Jour. 35(2):202-204.
- (9) \_\_\_\_\_, and Heinemann, B.  
1953. Techniques Used in the Babcock Test for Milks in the United States. Ill. Agr. Expt. Sta. Cir. 709, 28 pp.
- (10) Herrmann, L. F., Bryan, W. G., and Anderson, E. D.  
1954. Sampling Routines and the Accuracy of Patrons Butterfat Tests. U. S. Dept. Agr. Mktg. Res. Rpt. 66, 23 pp.
- (11) Hoover, S.R., Mucha, T. J., and Harvey, W. R.  
1958. A Comparison of Detergent Tests for Butterfat in Milk with Official Methods. Jour. Dairy Sci. 41(3):398-408.
- (12) Horwitz, M., Chairman and Editor.  
1955. Official Methods of Analysis of the Association of Official Agricultural Chemists. Ed. 8, Washington, D. C. 1008 pp. illus.
- (13) Lampert, L. M., Nelson, D. H., and Wilster, G. H.  
1952. The Procedure and Equipment for Determining the Fat in Milk by the Babcock Method. Report of the Committee of the American Dairy Science Association Appointed to



Standardize Methods for Conducting All Phases of the Babcock Test. W. Div. Amer. Dairy Sci. Assoc. Proc., pp. 41-49.

- (14) Marquardt, J. C.  
1950. 1950 Presents the Weigh Tank Problem. Amer. Butter and Cheese Rev. 12(2):18-20.
- (15) \_\_\_\_\_ and Durham, H. L.  
1932. Sampling Milk for Fat Test at Milk Plants. N. Y. (Geneva) Agr. Expt. Sta. Bul. 605.
- (16) Osgood, C. P.  
1950. Further Discussion of Country Plants and Butterfat Losses Encountered. State Insp. Dept. Agr. Augusta, Maine. Canadian Dairy and Ice Cream Jour. 29(8):70, 72, 74.
- (17) Preston, H. J.  
1954. Developing Butterfat Sampling and Testing Programs. Farmer Co-op. Serv. Bul. 5, 52 pp.
- (18) Ragsdale, H. L., Cook, E. W., Zimmerman, A. F., and Shupp, B. A.  
1955. Handling Butterfat Samples on Tank Trucks in Summer Temperatures. Amer. Milk Rev. 17(7):66,68 and 104.
- (19) Schwarzkopf, V.  
1950. Effect of Weigh Can Design on Butterfat Samples. Milk Dealer, 40:176-183.
- (20) Watson, P. D.  
1958. Effect of Variations in Fat and Temperature on the Surface Tensions of Various Milks. Jour. Dairy Sci. 41(12):1693-1698.
- (21) Wilster, G. H., and Robichaux, R. P.  
1940. Sampling, Preserving, and Testing Milk. Oreg. Agr. Expt. Sta. Bul. 383, 44 pp. illus.





Growth Through Agricultural Progress





