RELIABILITY OF SUBJECTIVE JUDGMENTS IN THE INSPECTION OF HARD RED WINTER WHEAT

by

CALVIN KELLY ADAMS

B. S., Kansas State University, 1958

A THESIS

submitted in partial fulfillment of the

requirements for the degree

MASTER OF SCIENCE

Department of Psychology

KANSAS STATE UNIVERSITY
OF AGRICULTURE AND APPLIED SCIENCE

1960

TABLE OF CONTENTS

INTRODUCTION

This paper reports a psychophysical investigation of subjective judgments made in the inspection of Hard Red Winter Wheat (HRWW), using a modification of the constant method in a field situation.

The United States Grain Standards Act as amended (United States Department of Agriculture 1941) specifies that all grain entering into interstate or foreign commerce which is sold, offered for sale, or consigned for sale by grade shall be inspected by a licensed grain inspector under the supervision of federal grain supervisors. The Act specifies how the inspection shall be conducted, what determinations shall be made, and what grades and sub-classes shall be assigned as a result of these determinations. The Act also sets forth a system whereby any grain owner not satisfied with the results of any inspection may call for a reinspection of the lot of grain, or appeal the inspection to the Federal Grain Supervisors. The purpose of the Act is to provide a common system and standardized procedures for evaluating the commercial value of grain.

Working under this Act, which is administered by the Agricultural Marketing Service of the United States Department of Agriculture, licensed grain inspectors inspect samples from millions of lots of grain each year. Some of the determinations which the inspectors must make, such as weight per bushel, moisture content, and dockage are rather routine, objective measurements, requiring little or no subjective judgment. The principle sources of error in these determinations are mechanical and sampling errors rather than "human" errors. Most of the determinations, however, require the inspector to identify certain qualitative factors in or characteristics of the sample, then to separate out and determine the weight of these elements of the sample. A case in

point is the evaluation of various forms of damage in wheat. Here the inspector must separate out by hand on the basis of color and morphological characteristics all of several types of damaged kernels. The damaged kernels are then weighed and the weight recorded as a percentage of the weight of the sample. The specific determinations which must be made vary from one grain to another.

In the inspection of hard red winter wheat in addition to the objective evaluations of weight per bushel, moisture content and dockage, and the subjective determination of damage described above, judgments are required in the determination of the percentages of wheats of other classes (WOOC), foreign material (FM), and dark, hard, and vitreous kernels (DHV). Each of these determinations require judgments based on visual cues of color and/or morphological characteristics. A final determination, the presence and nature of commercially objectionable foreign odors, requires a judgment based on olfactory experience. Thus, five of the eight factors which must be evaluated in the inspection of this one grain include important elements of subjective judgment by the grain inspectors.

The procedure followed in the determination of grade and subclass of a commercial lot of hard red winter wheat in the normal field situation is as follows: First, a representative sample of about 2 quarts is taken from the lot of grain, usually about 2000 bushels, by making 5 probes in the lot with a 5 foot double-tube compartment trier. The sample is then taken to the inspection point where it is first run through a Boerner divider to obtain 2 representative portions. Moisture content is then checked. After this, dockage is determined on approximately 1000 grams ($1\frac{1}{8}$ quarts) of the original sample. The weight per bushel is then determined using the dockage-free portion of the 1000 grams. Following this the inspector will make whatever determinations are indicated by a close appraisal of the entire sample given him. If he determines

heat and/or total damage or foreign material, he will divide the sample down to approximately 50 grams by running it through a Boerner divider a number of times. He then will accurately weigh out 50 grams from which he will make his determination. The procedure is the same for determining subclass (per cent DHV) and wheats of other classes (WOOC) except that for these determination he cuts the 1000 gram sample to 25 grams. He uses a different portion of the sample for each of the determinations made. He may also inspect the entire 1000 grams for odor, stones and cinders, insects, etc. if these determinations are needed. He then makes out a ticket specifying the results of inspection in terms of grade and subclass and indicates the percentage of the factor determining grade plus any other factors required by the grain standards as they apply to that sample.

The grade and subclass requirements for Hard Red Winter Wheat are given in Table 1.

There is good empirical as well as theoretical reason to believe that these judgments are not without error and unreliability. The judgments require fine discriminations on the basis of limited cues, and, in some cases, even require the setting of arbitrary cutting points on what are essentially continuous gradations. Of the more than three million lots inspected during the 1958 fiscal year, 75,000, or approximately 2.5 per cent were appealed to the federal supervisors.[1] Reports from members of the grain industry confirm these findings. One representative of a large milling company reported that a check at one inspection point indicated discrepancies in grade and/or subclassification between the reports of licensed inspectors and those of equally

---

[1]J. E. Elstner, Personal Communication, March 1959.

Table 1

Grade and Subclass Requirements for Hard Red Winter Wheat

Grade	Minimum Test Weight Per Bushel lbs.	Maximum Limits of				
		Damaged Kernels			Wheats of Other Classes	
		Total %	Heat-damaged Kernels %	Foreign Material %	Total %	Durum and/or Red Durum %
1*	60	2.0	0.1	0.5	5.0	0.5
2*	58	4.0	0.2	1.0	5.0	1.0
3*	56	7.0	0.5	2.0	10.0	2.0
4	54	10.0	1.0	3.0	10.0	10.0
5	51	15.0	3.0	5.0	10.0	10.0

Sample: Sample grade shall be wheat which does not meet the requirements for any of the grades from No. 1 to No. 5, inclusive; or which contains more than 15.5 per cent of moisture; or which contains stones; or which is musty, or sour, or heating; or which has any commercially objectionable foreign odor except of smut or garlic; or which contains a quantity of smut so great that any one or more of the grade requirements cannot be applied accurately; or which is otherwise of distinctly low quality.

*The wheat in grades No. 1 and No. 2 of this class may contain not more than 5.0 per cent and in grade No. 3 not more than 8.0 per cent of shrunken and broken kernels.

Sub-Classes: Class IV Hard Red Winter Wheat shall include all varieties of hard red winter wheat and may include not more than 10.0 per cent of wheats of other classes. This class shall be divided into the following three subclasses:

1. Dark Hard Winter Wheat. The subclass Dark Hard Winter Wheat shall be Hard Red Winter Wheat with 75 per cent or more of dark, hard, and vitreous kernels.

2. Hard Winter Wheat. The subclass Hard Winter Wheat shall be Hard Red Winter Wheat with 40 per cent or more but less than 75 per cent of dark, hard, and vitreous kernels.

3. Yellow Hard Winter Wheat. The subclass Yellow Hard Winter Wheat shall be Hard Red Winter Wheat with less than 40 per cent of dark, hard, and vitreous kernels.

qualified company inspectors on nearly 50 per cent of the lots inspected.[2] It is also interesting to note that in reviewing the pertinent literature, not one objective investigation of the reliability of these subjective judgments was found.

It has been reflected by some members of the grain industry that the inspectors' task has been made increasingly more difficult by the ever-increasing number of hybrid and crossbred varieties and sub-varieties being developed within each grain. Relatively distinct morphological characteristics have become less recognizable with the infusion of these new varieties.

There is good reason to believe, also, that the judgments made in grain inspection are subject to some of the same sources of error that have been identified in numerous psychophysical and psychometric investigations. A number of these have been summarized elsewhere by Guilford (1954) and Johnson (1955). Two important classes of variables which may affect these judgments are (1) adaptation effects, and (2) the physical conditions under which the judgments are made. Adaptation effects have been identified in the judgments of a variety of stimulus characteristics. Essentially, they refer to the effects of other stimuli in a series on the judgment of the Nth stimulus. Thus, it might be predicted that the judgment of the percentage of dark, hard, and vitreous (DHV) kernels in a sample with a relatively large proportion of such kernels may be affected by the contrast between that sample and the samples immediately preceding it. As to the effects of physical conditions of the environment in which the judgments are made, while it is recognized that the human observer is capable of maintaining relatively good perceptual constancy under varying conditions, it is also recognized that such constancy is not

---

[2] A. R. Baldwin, Personal Communication, May 1959.

perfect, and, in the absence of sufficient contextual cues, may be quite unstable, Therefore, even though a sample of wheat may be seen as "dark" under a variety of conditions of illumination, one may also anticipate some variability in judgment as a result of these conditions.

It was with such thoughts as these that the present study was begun. As an initial investigation in a projected series of studies on the judgments involved in grain inspection, its purpose was to objectively evaluate the inter- and intra-inspector reliability and accuracy of the subjective judgments currently being made by inspectors under existing field conditions. The scope of the study was limited by a number of practical considerations. First of all, since it included as subjects persons who are practicing grain inspectors at various points in Kansas and Missouri, both their time and the resources of the researchers were limiting factors. Complete inspection of a sample of wheat, for example, requires on the average about 20 minutes. Furthermore, it was concluded that it was more desirable to obtain a more adequate body of data on a limited number of variables than to attempt to generalize from limited data on a larger number of variables. For these reasons the experimental samples were made up of Hard Red Winter Wheat (HRWW), one class of one of the eleven commercial grains covered by the Grain Standards Act. HRWW constitutes a major portion of the grain produced in the Midwest. Furthermore, its inspection requirements include many of the problems of judgment involved in the inspection of other grains.

Judgments involved in four of the eight determinations made in the inspection of HRWW were selected for study: (1) percentage of dark, hard, and vitreous kernels (DHV), (2) percentage of wheats of other classes (WOOC), (3) percentage of foreign material (FM), and (4) percentage of damaged kernels (damage). The evaluation of judgments involved in the determination of

commercially objectionable foreign odors would require special apparatus and samples. This factor merits a separate study.

In addition to assessing the accuracy and reliability with which these variables are judged, the study was designed to provide preliminary evidence on the relationship of a limited number of situational and personal data variables to accuracy and consistency of judgments. To this end, data were collected on level of illumination, time of day each sample was worked, and age and length of service of the inspectors.

METHOD

Materials

Unlike many psychophysical studies in which judgments are made of the same (identical) physical stimulus or series of stimuli by a sample of judges, the judgments of grain samples present a special problem in that the inspection procedures essentially destroy the experimental sample for further use. The inspector separates out those kernels which he judges to be, in one instance, dark, hard, and vitreous, or in other instances, foreign material, wheats of other classes, or damaged kernels. He may, on occasion, bite or cut one or more kernels to confirm his judgment of vitreousness of carmelization (heat damage). In addition, small losses or changes in the sample as a result of the several handling and weighing operations may change the identity of the sample.

For these reasons, it was necessary to prepare equivalent samples for each inspector, rather than to use the same identical set of samples for all inspectors. While fairly homogeneous samples could possibly have been obtained by

dividing a well-mixed, well-blended master sample of known properties into as many subsamples as there were inspectors, such a procedure would not readily permit an evaluation of the variability among the subsamples. As a result, variability among inspectors' reports would have been confounded with variability inherent in the samples.

The alternative procedure was to prepare samples of known values of each of the factors under investigation, weighing each factor for each sample separately. This would insure equivalence within measurable tolerances of errors due to the weighing operations and to contamination within the component parts.

This latter procedure was followed. First, population lots of each component factor had to be prepared. For the DHV and Yellowberry (YB = Non-DHV) components, two commercial lots of HRWW were obtained, one of which contained less than 4 per cent YB, the other contained less than 35 per cent DHV. The first of these was picked as the population lot of DHV, the second as the population lot of YB. These population lots were then carefully and conservatively hand picked to minimize contamination of each by the other component. About 14,000 grams of each population were picked. They were then reinspected for contamination not removed by the hand picking. Twenty samples of 10 grams each were taken from both the DHV and YB populations and conservatively hand picked again, and the amount of contamination, by weight, recorded. Samples were drawn randomly in the following manner: Each population was first carefully blended in a Boerner divider. Then the population was divided into 20 approximately equal parts. Each part was then run through a divider separately and by careful weighing, a 10 gram sample was obtained from each of the divided portions. Hence, the 20 samples were drawn from throughout the total population. The mean amount of YB found in the 20 DHV samples was .115 grams (1.15 per cent)

with a standard deviation of .057 grams (.57 per cent). For the 20 YB samples the mean amount of DHV was .095 grams (.95 per cent), with a standard deviation of .057 grams (.57 per cent). With this information, 99 per cent confidence limits could be placed upon any programmed percentage of DHV. The length of the confidence interval was 1.46 grams for all the values of DHV programmed. Due to the unequal amounts of each population placed into each of the samples, the upper and lower limits were not equally distant from the programmed value of DHV, but varied progressively in accordance with the amount of each population programmed. The procedure used to obtain these confidence limits and a table of the resultant values of the confidence limits are given in Appendix A.

Certified samples of soft red winter wheat and of rye were acquired from the Agronomy Department of Kansas State University to be used as the populations for WOOC and FM, respectively. These represent common types of WOOC and FM. These populations were essentially free of contamination.

Preparation of a population of damaged wheat presented some difficult problems. There are a number of kinds of damage, including heat damage, sick or black-germ, sprout, frost, mold, insect, and weather damage, which the inspector must identify as "damaged kernels". Only heat damaged kernels, however, must be identified and weighed separately. All other types of damage are reported only as total damage.

To simplify the weighing and programming of the samples, three types of damage, sick or black germ, sprout, and heat damage, were selected. A population of sprout damaged wheat was prepared by placing a portion of the yellow berry population in a germinating oven until it was evident that the majority of kernels were sprouted. This preparation was then conservatively hand picked to minimize contamination with non-sprouted kernels. Following the hand picking, ten samples of ten grams each were drawn randomly from this population.

These samples were again conservatively hand picked for non-sprouted kernels and contamination, and the values, by weight, recorded. The mean amount of non-sprout for the 10 samples was 0.36 grams (3.6 per cent), with standard deviation equal to .192 grams (1.92 per cent). A heat damage population was obtained by placing another portion of the YB population in a drying oven at a temperature of 100° C for 4 hours, it was then carefully checked for purity. It was judged to be 100 per cent heat damaged by two subject matter experts.[3]

A population of sick or black germ damaged wheat was prepared by placing a third portion of the yellow berry population in a drying oven at 50° C for 4 days. This preparation was also judged to be 100 per cent sick damaged.

Heat and sprout damage populations were blended in a 1 to 5 ratio and constitute the heat-sprout population. The heat and sprout were not kept separate for two reasons. One, it would have added another weighing operation, and two, the primary interest was in the black germ damage, and the other forms of damage were added mainly to give the samples face validity.

With these six populations at hand, essentially equivalent sets of samples could be programmed with each sample within the set having different and known values of each of the components. For practical considerations, the number of samples in each set was limited to 16. Values in percentages of the total sample weight from each population were determined by first establishing realistic limits for each component, that is, the range within which values of each factor would normally fall, and then choosing values within these limits by means of a table of random numbers with the exception that values were arbitrarily selected at or near some of the crucial values for the determination of grade or subclass. After the distribution of values for each factor had been

---

[3]Professors Ernest Mader and Howard Wilkins, Department of Agronomy, Kansas State University.

chosen, the values of the four factors were combined randomly to make up the 16 master samples, with the restriction that the percentages of the four factors could not total more than 100 per cent. The residual was made up from the YB population. The values programmed into the 16 master samples are presented in Table 2.

Table 2

Composition of the 16 Experimental Samples

Sample No.	DHV %	WOOC %	FM %	Per Cent Damage			Grade	Sub-Class
				Sick %	Sprout & Heat %	Total %		
1	64	8.0	6.0	6.75	8.25	15.0	6	Hard Winter
2	16	66.0	1.0	7.80	5.20	13.0	5	Yellow Hard
3	19	9.0	2.5	0.2	1.80	2.0	4	Yellow Hard
4	94	0.0	0.0	0.75	2.25	3.0	2	Dark Hard
5	89	0.0	5.0	0.3	0.7	1.0	5 or 6	Dark Hard
6	78	5.0	5.0	0.0	7.0	7.0	5 or 6	Dark Hard
7	23	4.0	0.5	2.0	2.0	4.0	2 or 3	Yellow Hard
8	57	3.0	3.0	0.60	3.40	4.0	4 or 5	Hard Winter
9	51	7.0	4.0	0.25	4.75	5.0	5	Hard Winter
10	28	8.0	3.0	3.20	12.8	16.0	6	Yellow Hard
11	33	2.0	2.0	6.50	3.5	10.0	4 or 5	Yellow Hard
12	74	3.0	2.0	3.20	4.8	8.0	4	Hard Winter
13	42	1.0	1.0	0.70	1.3	2.0	2 or 3	Hard Winter
14	37	10.0	1.5	9.0	3.0	12.0	5 or 6	Yellow Hard
15	83	11.0	0.5	0.0	0.0	0.0	6	Dark Hard
16	12	10.0	0.0	3.3	2.7	6.0	3 or 6	Yellow Hard

Seventy samples of each of the 16 master samples were prepared. Each sample required six weighing operations (DHV, YB, FM, WOOC, Heat and Sprout Damage, and Sick or Black Germ Damage). As a check on the reliability of weighing, five of the 70 samples were re-weighed after each weighing operation as each of the

16 master samples was prepared. Tolerance limits were set at one per cent error for any stage of the weighing process. Thus, if in the Nth sample the first two weighings should give a total weight of 10 grams, anything greater than ± .1 gram error would be refused and that sample thrown out. Total weight after all operations was, therefore, within 1 per cent of 25 grams, (± .25 grams). All weighing was done by experienced personnel using torsion balances accurate to 1/100 gram.

Each 25 gram sample was placed in a paper envelope and sealed. Envelopes were given code numbers to identify each of the 16 different master samples, and the inspector to whom they were administered. One envelope from each of the 16 constituted a sample set. The order of the samples in a set was randomly selected from a table of random orders of 16 events, a different order being obtained for each set.

Subjects

A total of 40 licensed state grain inspectors practicing at nine Kansas and one Missouri inspection stations served as experimental subjects for this study. These inspectors range in age from 22 to 70 years, with a median age of 54. In years of experience, they range from 0 to 40, the median being 16 years.

Procedure

To obtain a sample of from 40 to 50 practicing licensed grain inspectors, the directors of the Kansas and Missouri Grain Inspection Departments were contacted and the purpose and nature of the study explained. The whole-hearted

cooperation of both state inspection departments was obtained.

The Directors of the Kansas and Missouri Grain Inspection Departments notified their inspectors of the proposed project about two weeks before data collection was to begin. Cooperation in the project was requested, and the approximate dates on which members of the research team would arrive were announced. Individual stations were contacted by the researchers a few days before the station was to be visited. All sample sets were administered by the writer and three graduate student assistants.

Between June 2 and June 12, 1959, complete sample sets were inspected by 26 inspectors in Kansas and Missouri. Three additional inspectors completed parts of the sample sets. Approximately two months later, between August 3 and August 14 sample sets were again completed by 21 of the 29 inspectors who had worked the samples in June, and by 11 inspectors who had not previously worked the samples. Hence, there were three samples of inspectors; those who worked the sample sets in June only (N = 8), in August only (N = 11), and in both June and August (N = 21).

The following instructions were read to each inspector before he began to work the sample set:

> There are 16 25-gram samples for you to inspect. These inspections will not be routine in that the information I want you to give me is more detailed than that which you normally give. Otherwise, these inspections will be routine in that they will be on the factors you normally look for. This is the procedure I would like you to use in inspecting these samples:
>
> 1. Inspect each 25-gram sample in the order in whichy they are numbered. Each inspector will be identified by number only and this number will be the first two digits of the number on each envelope. The last four digits are the sample number - so inspect the 16 samples in the order of these numbers, working from low to high.
>
> 2. Inspect each 25-gram sample for these four factors: per cent DHV, per cent WOOC, per cent FM, and per cent Damage. Record the percentages as you find them for each factor on the appropriate line on the envelope from which the sample was taken. Figure the per cent to the nearest 1/16 per cent. You may determine these factors in whatever order you desire,

in other words, work as you usually do, as far as the order of determintions is concerned.

3. Regard all samples as fulfilling the requirements of grade one as far as moisture content, weight per bushel, dockage, and odor are concerned.

4. When determining damage, I would like to have you determine total damage, and also the per cent of each type of damage in the sample. Record those types of damage not listed on the envelope on the lines allowed.

5. After you have completed the four determinations, give the grade and subclass into which those four factors would place that sample.

6. Work at your own pace, and do the best job possible. When you start each new sample, indicate this to me as I wish to take a light meter reading for each sample.

7. Please replace all of the sample in the packet when completed.

8. All 16 samples each inspector will receive are different, and the samples each inspector gets are different from those received by any other inspector, as a matter of fact, these samples here are part of 1600 samples which we are having evaluated.

9. The purpose of this study is threefold: (a) to investigate the "human element" or the tolerance limits within which licensed inspectors can be expected to operate, (b) to determine the accuracy under different combinations of samples, and (c) to obtain a set of carefully inspected samples for classroom work.

Time of day and illumination level were recorded at the beginning of each sample in the set. Illumination readings were taken with Argent Hyper VII light meters. Several omissions in these data will be noted. These resulted from conflicting demands on the researcher's time while administering more than one set at a time, and from a faulty light meter. Discrepancies from standard inspection procedures were also noted; for example: some inspectors insisted that they never worked with more than a 10 gram sample and proceeded to divide the samples so as to obtain 10 grams.

RESULTS

Inter-Inspector Reliability

The extent to which different independent inspectors, working under similar physical conditions, are able to agree in the estimation of equivalent values of each of the four subjective factors investigated is here referred to as inter-inspector reliability. The objective is to determine the variability, direction, and magnitude of any disagreements in estimates exhibited in the 29 sample sets evaluated in June and the 11 sample sets evaluated in August by non-repeat inspectors.

To do this the mean, standard deviation, and range of the inspectors estimates was obtained by summing over the 40 inspectors estimates for each value of each factor separately.

Sub-Class

The results of this procedure for the values of DHV programmed into the 16 samples are presented in Figure 1. Each bar on the graph represents the combined estimates of 40 different inspectors, except where occasional omissions were made. The white portion of each bar represents the range within which the true per cent of DHV programmed is expected to fall 99 times out of 100. The pip on top of each bar represents the mean of the inspectors estimates. The double width portion of the bar includes one standard deviation on either side of the mean on the inspectors estimates, and the length of the single width bar represents the total range of the estimates given by the inspectors for each programmed value of DHV. The two vertical lines on the graph represent the cutting points for the three subclasses (40 per cent and 75 per cent DHV).

Inspection of the data of Figure 1 indicates several important results.

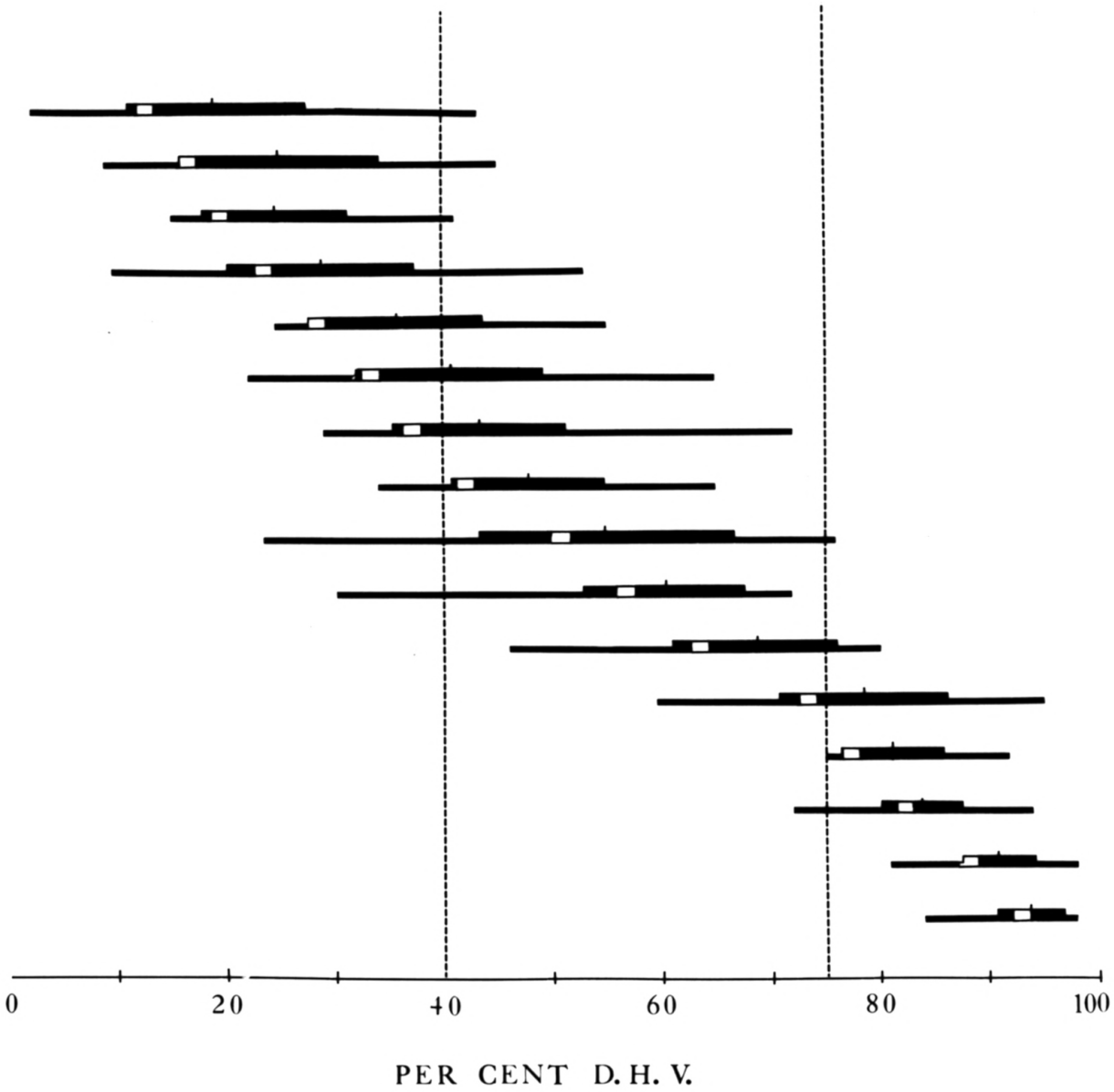

Fig. 1. Mean, standard deviation, and range of inspector's estimates for each of 16 values of DHV (N = 40 inspectors).

Note.--White portion of each bar represents 99 per cent confidence interval for programmed value of DHV.

First, these inspectors did disagree in their estimates of all of the 16 programmed values of DHV. It may be noted that for all 16 values of DHV the mean of the inspectors estimates falls outside of the range of expected true values. It may be noted further that the means of the inspectors estimates are always above the true value and that the magnitude of this over estimation increases as the programmed value of DHV gets smaller. It should be noted also that at least some of the inspectors estimates placed the sample in the wrong subclass for all except three of the programmed values of DHV.

A summary of the inspectors estimates for DHV is presented in Figure 2. The three curves, extrapolated from data on the 16 samples, are estimates of the proportion of times a sample with a given true value of DHV will be placed in each of the three subclasses. To illustrate, the curve fitted through the open circle data points shows the proportion of times a sample with any given value of DHV will be placed in the lowest subclass (Yellow Hard W.W.). If, for example, a sample had 30 per cent DHV it would be placed in the lowest subclass by about 70 inspectors out of 100, the other 30 inspectors would place the sample in the middle subclass. Figure 2 reveals two results: First, the likelihood of a sample being placed in the proper subclass decreases as the percentage of DHV contained approaches a subclass cutting point, and second, inspectors tend to overestimate the per cent DHV, as shown by the points of intersection of the curves. These points show the value of DHV at which a sample has an equal chance of being placed in either of two subclasses. If the inspectors showed variable error but no constant errors of estimation, the curves would intersect exactly at the two subclass cutting points, but in fact they intersect at values of DHV below the cutting points. This difference represents the average amount by which the inspectors overestimate the per cent DHV.

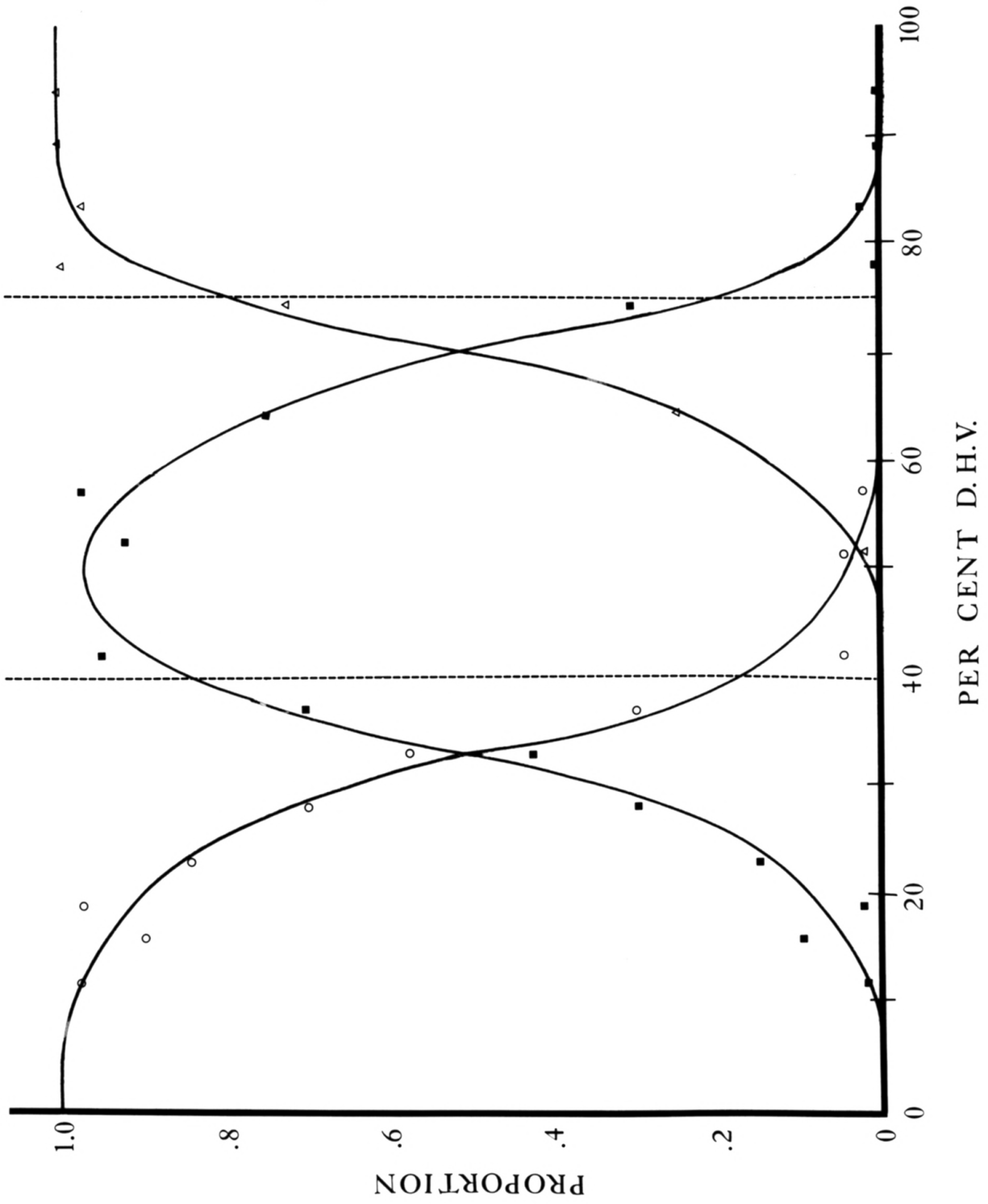

Fig. 2. Summary curves extrapolated from observed data points giving proportion of time a sample with a given percentage of DHV will be placed in each of the three sub-classes (circles - Yellow Hard, squares - Hard Winter, and triangles - Dark Hard).

The results of the inspector estimates of DHV, and the three other factors are presented in tabular form in Table 16 in Appendix B.

Grading Factors

Wheats of Other Classes. The results of the inspectors estimates of WOOC are shown in Figure 3. This figure is essentially the same as that for DHV with two exceptions. First, the range of values covered is much smaller, and second, the true programmed value of WOOC is represented by the inverted triangle above each bar. The two vertical lines represent the two grade cutting points. It is interesting to note the rather marked amount of variability in the estimates of all values in relation to the grain standards. There is no consistent trend in the mean of the inspectors estimates over the different values programmed they tend to overestimate the per cent WOOC but not consistently.

There are two other points of interest in Figure 3. First, at least some inspectors found some WOOC in the two samples into which actually no WOOC was programmed, and second, all but three of the samples were placed in two or more grades by at least some inspectors.

The lack of any consistent trend in the inspectors mean estimates for the different amounts of WOOC was analyzed further. Inspection of the data suggested that the mean error of the inspectors estimates of WOOC appeared to be inversely related to the amount of DHV programmed into the samples. To test this possibility a partial correlation coefficient was calculated to evaluate the degree of relationship between the programmed value of DHV and the inspectors mean error of estimate on each sample over the 16 different samples, with the programmed per cent of WOOC partialled out.

To evaluate this relationship, three first-order correlations were obtained: programmed per cent DHV x inspectors mean algebraic error on WOOC over the 16 samples ($r_{12} = -.685$); programmed per cent DHV x programmed per cent WOOC

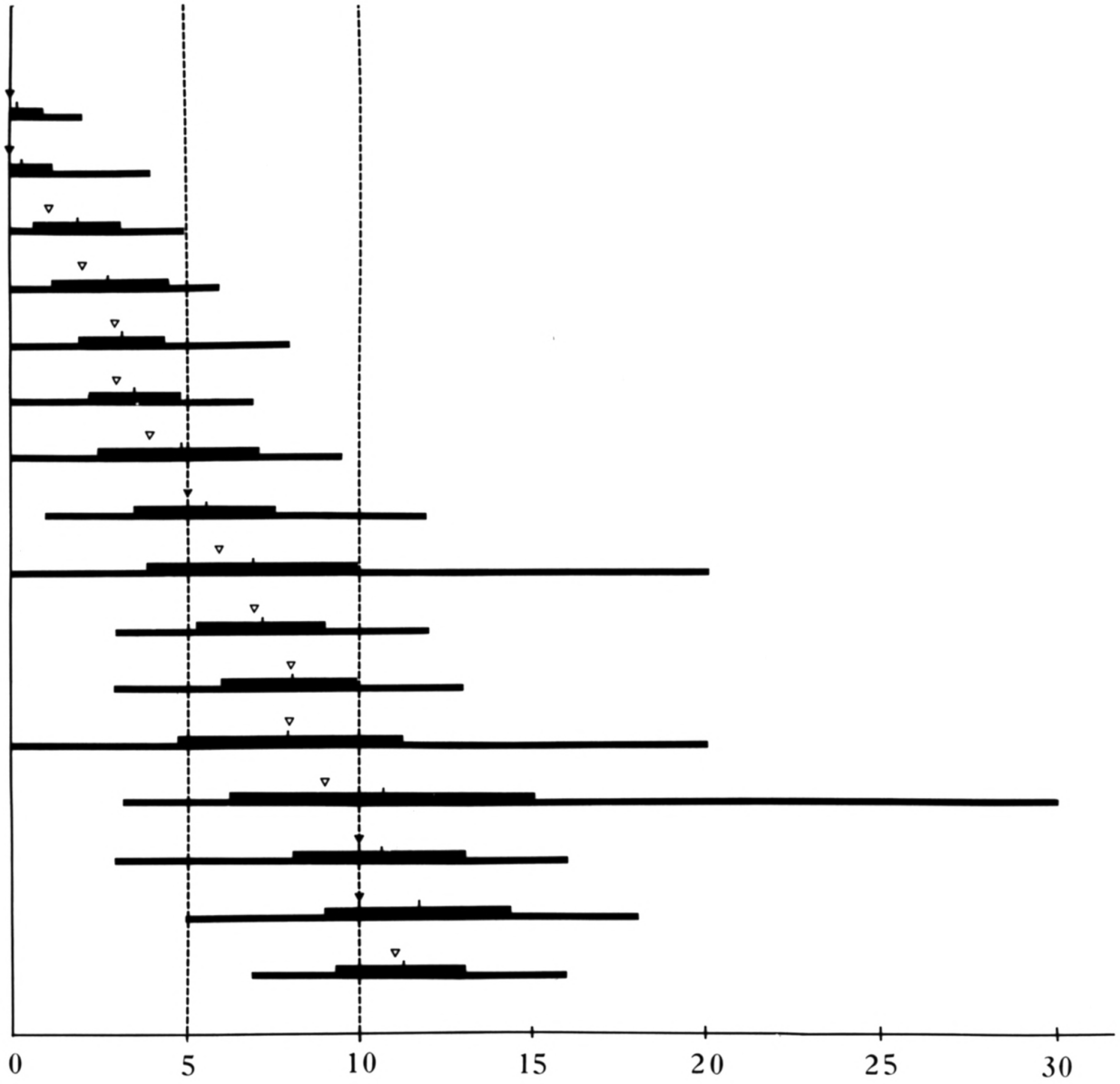

Fig. 3. Mean, standard deviation, and range of inspectors' estimates for each of 16 values of WOOC (N = 40 inspectors).

Note.--Inverted triangle above each bar represents programmed percentage of WOOC.

over the 16 samples (r_{13} = -.393); and programmed per cent WOOC x inspectors mean algebraic error on WOOC over the 16 samples (r_{23} = .268). These coefficients were then used in the standard formula for obtaining a partial correlation, which is:

$$r_{12 \cdot 3} = \frac{r_{12} - (r_{13}\, r_{23})}{\sqrt{(1 - r_{13})^2\,(1 - r_{23})^2}}$$

The obtained value of $r_{12 \cdot 3}$ = -.65 with 13 d.f. was significant beyond the .01 level.

These results show that the inspectors tended to pick YB as WOOC and the amount of this confusion of factors increased as the amount of YB in the sample increased, that is, as the amount of DHV decreased.

This relationship was further strengthened when a second partial correlation was run evaluating the relationship between inspectors mean algebraic error on WOOC and programmed per cent WOOC, with programmed per cent DHV partialled out ($r_{23 \cdot 1}$ = .002).

Foreign Material. Inspection of Figure 4 will show the results of the inspectors estimates of FM. This graph is similar to that for WOOC except for the more restricted range of values and the added grade cutting points. It should be noted again that there was rather pronounced variability in the inspectors estimates relative to the range of the standards. It should be noted also that on this factor the inspectors mean estimates tend to be consistently under the programmed value of FM and that this tendency increases as the amount of FM programmed increases. Furthermore, with the added number of grade cutting points, only the sample with 0 per cent FM is not placed in two or more grades by the inspectors estimates.

Damage. It was felt when this investigation was initiated that the inspection of damage would probably prove to be the most difficult for the

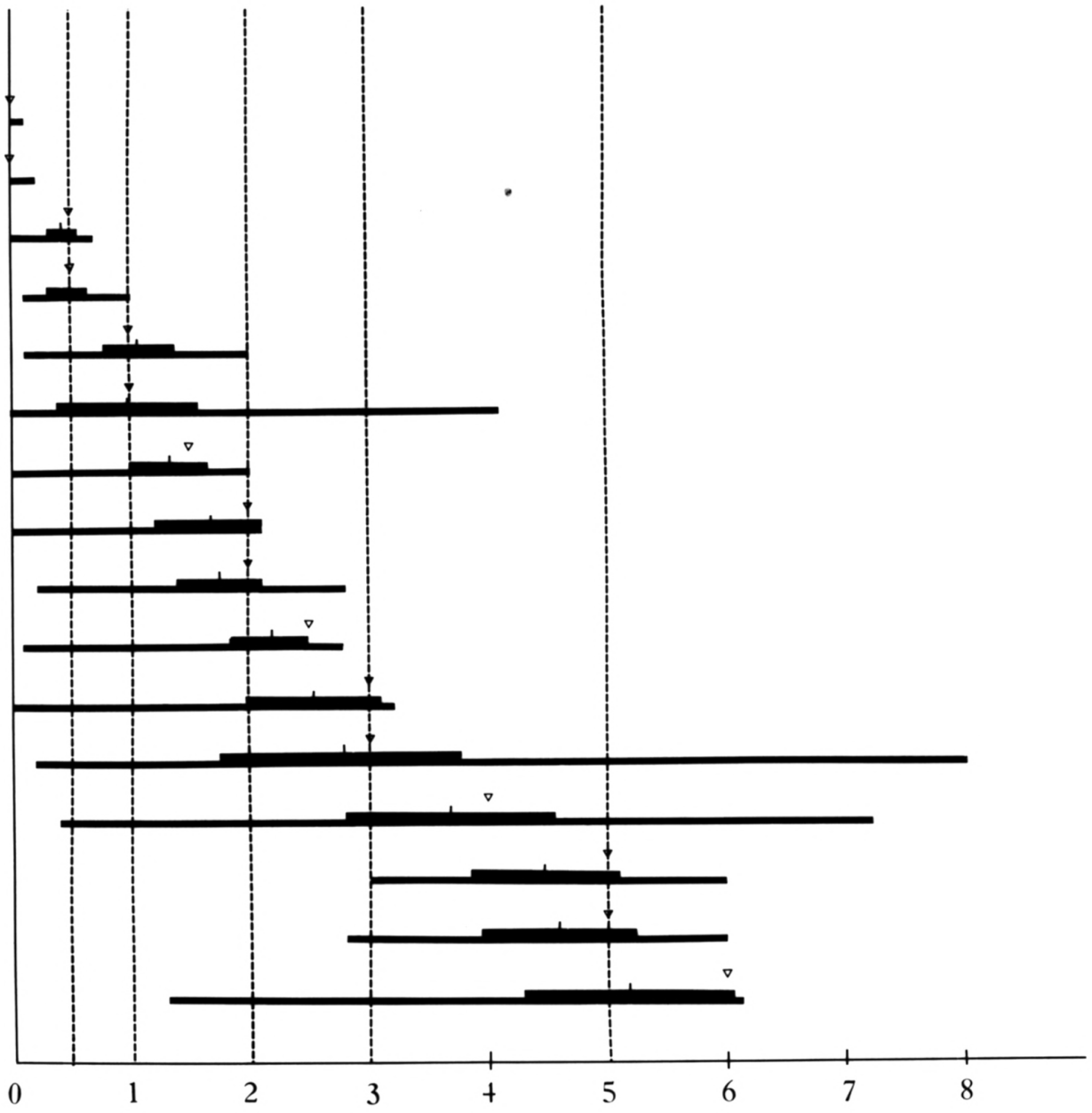

Fig. 4. Mean, standard deviation, and range of inspectors' estimates for each of 16 values of FM (N = 40 inspectors).

Note.--Inverted triangle above each bar represents programmed percentage of FM.

inspectors. The results shown in Figure 5 support this expectation. The results shown are for estimates of total damage. Here again the inspectors show considerable variability in their estimates relative to the standards. There is also a tendency to underestimate the amount of damage in any sample and this tendency for underestimation grows stronger as the amount of damage programmed increases, but it is not entirely consistent. Possible causes of this lack of consistency in the inspectors mean estimate were investigated. When a partial correlation was run to evaluate the effect of the amount of sick or black germ damage in the sample on the error of the inspectors mean estimates of total damage, a marked relationship was found. Again three first order correlations were first obtained: Per cent sick damage programmed x inspectors mean algebraic error of total damage ($r_{12} = -.955$); per cent total damage programmed x inspectors mean algebraic error for total damage ($r_{23} = -.867$) and per cent total damage programmed x per cent sick damage programmed ($r_{13} = .796$). Using the same formula for obtaining a partial correlation as before, the resultant partial correlation $r_{12 \cdot 3} = -.880$ with 13 d.f. is significant at beyond the .01 level. This partial correlation indicates the degree of relationship between programmed per cent sick damage and the inspectors mean estimate of total damage with per cent total damage programmed partialled out. This means then, that as the amount of sick damage programmed increased, the inspectors tended to underestimate the percentage of total damage more and more.

A second partial correlation was also run ($r_{23 \cdot 1} = -.598$). This indicates that not all the error on estimation of damage was due to sick or black germ damage, i.e. the inspectors were also progressively underestimating the value of the sprout and heat damage, but not to as marked an extent.

Errors in Grade and Sub-Class

The previous analysis of inter-inspector reliability, while giving a

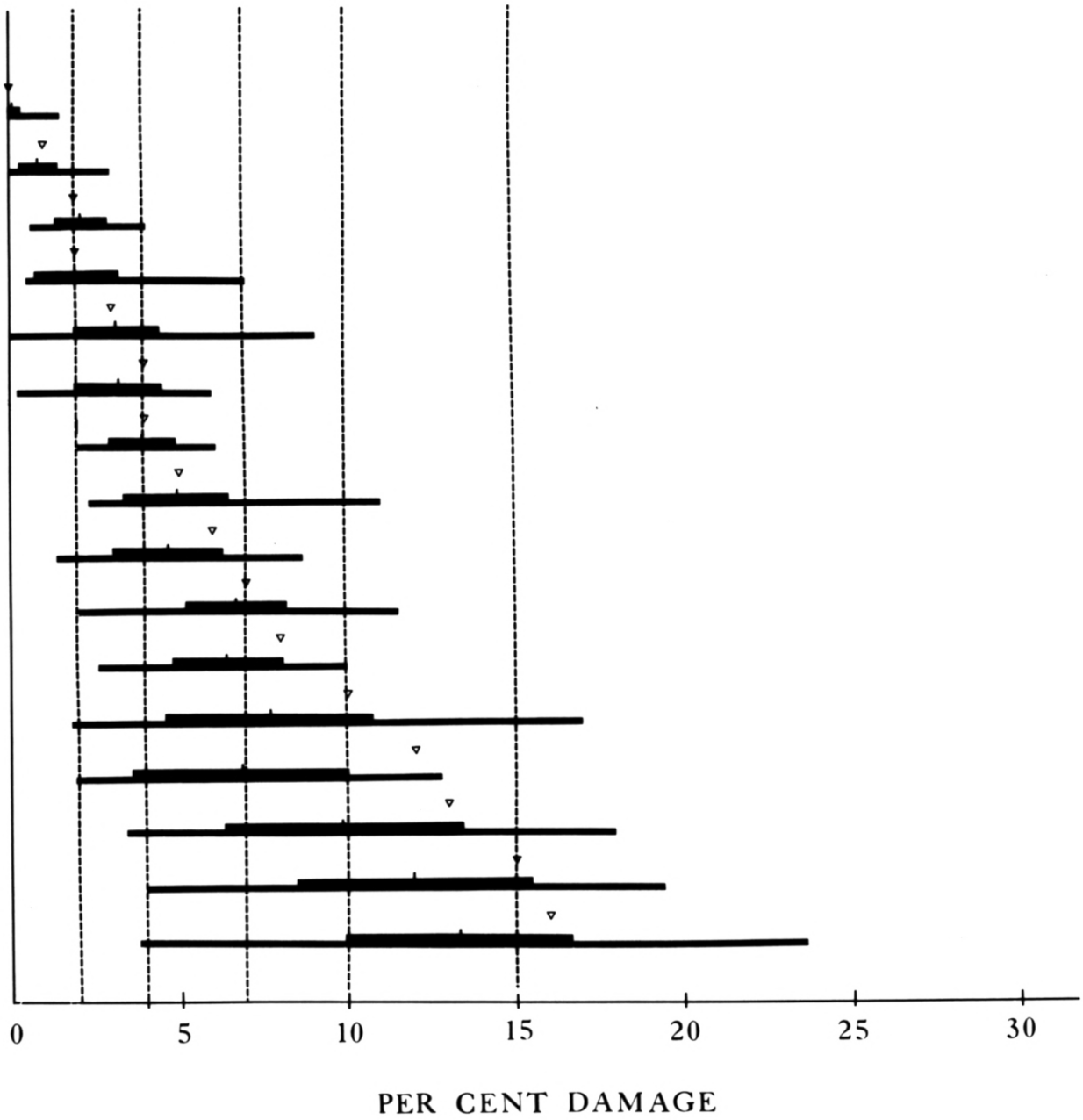

Fig. 5. Mean, standard deviation, and range of inspection estimates of each of 16 values of Damage (N = 40 inspectors).

Note.--Inverted triangle above each bar represents programmed percentage of Damage.

rather precise description of the results, gave no particular indication of how well the inspectors agreed on the grade or subclass of the different samples. As far as the reliability of the inspectors in relation to the grain standard is concerned, placing the sample in the proper grade and subclass is the critical test. Only when the estimate of a particular factor is sufficiently in error to place that factor in another grade or subclass does it affect their reliability in relation to the grain standards.

Sub-Class. The total number of errors, and the number of errors of subclass for each of the values of DHV may be seen by inspection of the first three columns of Table 3. The first column gives the percentage of DHV in each sample, the second column gives the number of inspectors whose estimate of that value of DHV resulted in the placement of that sample in the wrong subclass. The third column gives the number of inspectors who gave estimates for each value of DHV. At the bottom of these columns is the percentage of all the estimates that were in error, i.e., the number of estimates in error divided by the total number of estimates. The most interesting thing about these errors is their differential frequencies for the different values of DHV. Going from low to high values of DHV, as a subclass cutting point is approached, the number of errors increases, reaching their highest frequency for 37 and 74 per cent DHV. Once the cutting point is passed, errors drop sharply. This indicates, as before, that the inspectors were tending to overestimate the percentage of DHV and that this overestimation did result in a number of errors in subclass.

Grading Factors. The next three columns of Table 3 give the errors in grade for WOOC. The columns have same headings as the subclass tabulations with one exception. Here there are values that fall exactly on grade cutting points. In those instances where the value is on a cutting point between two

grades, the grade for that factor is either the grade it would normally be graded into, or the grade above it. For example, a sample value of WOOC programmed at five per cent WOOC would normally fall in Grades 1 and 2, but in this analysis any estimate up to and including 10.0 per cent WOOC is not considered as resulting in an error in grade. The reason for this is that the weighing operations for all of the values for the different factors were not without error. In each case there was some possibility that somewhat more than the specified amount was programmed. The maximum error, however, was less than 0.1 grams. This of course results in a more conservative estimate of the number of errors in grade made by the inspectors.

The results for WOOC are much the same as those for subclass, except that here errors in grade are made about equally often both above and below the cutting points, thus showing much less directionality in the errors made than was the case with subclass errors.

The FM portion of Table 3 has one feature not encountered with the two previous factors. Here there are five grade cutting points where with DHV and WOOC there were only two. As a result, errors occurred which changed the value by two or more grades. These errors are described in the additional column as multiple grade errors. As before, errors in either direction (raising or lowering the grade) are counted together. It will be noted that the percentage of all errors of grade are lower here than for previously described factors, even though there was considerable variability in the estimates of FM as shown by Figure 4. One reason for this relatively low percentage of errors is the large number of values (nine) which fall on cutting points and thus give a large range of possible estimates that result in no error in grade as defined in this table. It is interesting to note, however, that errors were still made on these values, giving some indication of the variability of the inspectors estimates.

Table 3

Summary of Errors in Grade and Subclass for Initial Inspectors

Master Sample No.	Errors in Subclass			Errors in Grade for Each Factor Independently										Errors in Composite Grade of Sample				
				WOOC			FM				Damage							
	Per Cent	Errors	N	Per Cent	Errors	N	Per Cent	Errors	Multiple Errors	N	Per Cent	Errors	Multiple Errors	N	Errors	Multiple Errors	N	Grade[b]	Grade Factor & Value
1	64	10	38	8	5	39	6	14	1	39	15[a]	6	6	39	13	0	39	6	FM (6%)
2	16	4	39	6	11	39	1[a]	6	0	39	13[a]	7	10	39	9	9	39	5	Damage (13%)
3	19	1	37	9	19	39	2.5	6	4	39	2[a]	0	0	39	6	16	39	4	FM (2.5%)
4	94	0	40	0	0	40	0	0	0	40	3	11	1	40	11	1	40	2	Damage (3%)
5	89	0	37	0[a]	0	37	5[a]	2	0	37	1[a]	1	0	37	2	0	37	5 or 6[c]	FM (5%)
6	78	0	39	5[a]	1	39	5[a]	2	0	39	7[a]	2	1	39	2	0	39	5 or 6	FM (5%)
7	23	5	39	4	9	39	0.5[a]	0	0	31	4[a]	9	0	39	5	0	31	2 or 3	Damage (4%)
8	57	1	38	3	3	38	3[a]	8	1	38	4[a]	1	0	38	9	0	38	4 or 5	FM (3%)
9	51	3	39	7	7	39	4	4	1	39	5	12	1	39	5	1	39	5	FM (4%)
10	28	12	38	8	11	40	3[a]	6	1	40	16	23	6	40	19	5	40	6	Damage (16%)
11	33	18	40	2	2	40	2[a]	1	1	40	10[a]	14	6	40	20	0	40	4 or 5	Damage (10%)
12	74	29	38	3	2	38	2[a]	1	2	38	8[a]	23	3	38	26	0	38	4	Damage (8%)
13	42	2	39	1[a]	0	39	1[a]	5	0	39	2[a]	2	0	39	4	0	39	2 or 3	FM (1%)
14	37	28	37	10[a]	2	39	1.5	5	1	39	12	9	23	38	3	11	38	5 or 6	WOOC (10%) & Damage (12%)
15	83	1	36	11[a]	13	39	0.5[a]	0	0	39	0	0	0	39	0	13	39	6	WOOC (11%)
16	12	1	32	10[a]	1	38	0	0	0	37	6	15	2	37	3	0	37	3 or 6	WOOC (10%)
Σ		115	606		86	622		60	12	613		135	59	620	137	56	612		
%		19.0			13.8			9.8	2.0			21.8	9.5		22.3	9.1			

[a]Values which fall on grade cutting points.

[b]Grade of Sample is lowest grade (numerically highest) received by any grading factor.

[c]When value of grade-determining factor falls on a grade cutting point, grade for that sample is considered correct if estimate fell in normally proper grade, or the next grade above it.

The results for damage clearly indicate the difficulty with which this determination is made. It will also be noted that the percentage of multiple errors is greatest for this factor. Especially of interest is the large number of multiple errors for the 12 per cent value of damage. Reference to Figure 5 will show that the distribution of estimates for this value of damage covers a considerable range, and the inspectors mean estimate is considerably below the programmed value. Also, by reference to Table 2, it will be seen that sick or black germ damage constituted three-fourths (9 per cent) of the total damage. This again indicates the difficulty the inspectors had in evaluating sick damage correctly.

Composite Grade and Subclass of Sample. So far, the percentage of errors in grade or subclass have been presented separately for each factor. In the field situation when an inspector evaluates a sample of grain, the grade of that sample is the lowest (numerically highest) grade into which any grading factor of the sample falls. For example, if a sample is free of FM and WOOC but has eight per cent total damage, the grade of that sample is Number 4 (provided of course, it meets the requirements for the other grading factors not investigated here). Subclass is independent of grading factors and depends only upon the percentage of DHV kernels. Because of this, it was of interest to determine the number of samples inspected that were placed in the proper grade and subclass by the inspectors. These results are shown in Table 4. In column 2 of this table, an error of one grade and/or one subclass on any sample is counted as one error. It must be remembered that here the error may occur in <u>any</u> of the factors investigated, as the final evaluation of the sample as a whole is determined by both the percentage of DHV <u>and</u> the grading factor or factors which fall into the lowest grade. Due to this requirement two columns are added to the table. Column 5 gives the grade or grades into which the

programmed values of the four factors place each master sample. The subclass is indicated in column 8 by name.

It will be noted that some of the samples may properly be placed in two grades. This is the result of the determining value for grade falling on a grade cutting point. In the case of sample 14, a combination of values of factors further complicates the picture. To clarify, the case of sample No. 14, the 10 per cent value of WOOC would place the sample in Grades 3 or 6 without error. The 12 per cent value of damage, however, places an added restriction since if the value of WOOC is estimated at 10.0 per cent or less, damage then becomes the factor determining grade, but if WOOC is estimated at 10.1 per cent or more then WOOC is the proper grade-determining factor and no error in grade results, as was previously discussed. It will also be noted that the samples do not fall into all grades (no Grade No. 1) or equally often into the grades covered. This is the result of the random manner in which values of the factors were combined in making up the master samples. The samples do fall fairly equally into the three subclasses.

The results shown in the table are not unexpected in light of what had been disclosed before. They do indicate, as before, that the samples in which damage is the grading factor tend to be the most difficult for the inspector to grade properly.

Generality of Inspector Estimates

It would be of considerable value to obtain some information about the generality of an inspector's ability to make accurate and reliable estimates for the different factors. It would be interesting to know, for instance, that if any inspector tends to underestimate the percentage of damage that he also tends to underestimate or overestimate the values of the other factors. If this were so, it would then be possible to discuss or describe the behavior of

Table 4

Summary of Errors in Grade and/or Sub-Class of Composite Sample

Master Sample No.	Errors in Grade and/or Sub-Class	Multiple Errors in Grade and/or Sub-Class	n	Grade[a]	Grade Determining Factor & Value		Sub-Class	DHV %
1	18	0	38	6	FM	(6%)	Hard Winter	(64)
2	10	9	39	5	Damage	(13%)	Yellow Hard	(16)
3	7	15	37	4	FM	(2.5%)	Yellow Hard	(19)
4	11	1	40	2	Damage	(3%)	Dark Hard	(94)
5	2	0	37	5 or 6[b]	FM	(5%)	Dark Hard	(89)
6	1	0	39	5 or 6	FM	(5%)	Dark Hard	(78)
7	5	3	31	2 or 3	Damage	(4%)	Yellow Hard	(23)
8	10	0	38	4 or 5	FM	(3%)	Hard Winter	(57)
9	6	2	39	5	FM	(4%)	Hard Winter	(51)
10	23	5	38	6	Damage	(16%)	Yellow Hard	(28)
11	30	0	40	4 or 5	Damage	(10%)	Yellow Hard	(33)
12	33	0	38	4	Damage	(8%)	Hard Winter	(74)
13	6	0	39	2 or 3	FM	(1%)	Hard Winter	(42)
14	18	11	37	5 or 6	Damage WOOC	(12%) (10%)	Yellow Hard	(37)
15	1	13	36	6	WOOC	(11%)	Dark Hard	(83)
16	4	0	32	3 or 6	WOOC	(10%)	Yellow Hard	(12)
Σ	185	59	606					
%	30.5	9.7						

[a]Grade of sample is lowest grade received by any grading factor.

[b]When value of determining factor fell on a grade cutting point, the grade of that sample was considered correct if estimate fell in either the normally correct grade, or the next grade above it.

inspectors in regard to the entire inspection procedure, rather than restricting any description to one particular factor at a time.

To evaluate the nature and degree of the relationship of the inspectors estimates on the four factors investigated, three indicies of each inspectors estimates were obtained. (1) The mean algebraic error of each inspectors estimates for each factor, (2) The mean absolute error of each inspectors estimates for each factor, and (3) The standard deviation of each inspectors estimates for each factor. If each of these measures for all of the 40 inspectors estimates is intercorrelated with the same measure for the other three factors, the question of the generality of the inspectors estimates can be at least partially answered.

Mean Algebraic Error. Inspection of Table 5 indicates that there is a moderate negative relationship between magnitude and direction of errors for DHV, and direction and magnitude of errors for FM and Damage. There appears to be no similar strong relationship between any of the other factors. This indicates that the degree to which each individual inspector tended to overestimate the percentage of DHV was related to the extent to which the same inspectors tended to underestimate the percentages of FM and Damage. This tendency was moderately strong for FM and somewhat weaker for Damage.

Mean Absolute Error. The general magnitude of relationship between the four factors is much stronger for amount of error, than was found for mean algebraic error. There appears to be a moderately strong relationship between amount of error made by the individual inspector on all of the factors except WOOC. The relationship between amount of error on WOOC and the other three factors, although considerably weaker, does indicate some slight tendency for relationship. These coefficients indicate that the amount of error per se made by an inspector on any of the factors except WOOC is to a moderate extent

Table 5

Summary of Inter-Correlations Between Estimates of the Four Factors Investigated

Mean Algebraic Error N = 40			
	WOOC	FM	Damage
DHV	-.245	-.593**	-.377*
WOOC		.093	.192
FM			.042

Mean Absolute Error N = 40			
	WOOC	FM	Damage
DHV	.258	.623**	.534**
WOOC		.331*	.239
FM			.502**

Standard Deviations of Estimates N = 40			
	WOOC	FM	Damage
DHV	.192	.164	.157
WOOC		.310	.392*
FM			.350*

*Significant at .05 level ($r_{.05}$ = .312)

**Significant at .01 level ($r_{.01}$ = .403)

related to the amount of error made on two of the other three factors, and to a slight extent to amount of error on WOOC.

Standard Deviation of Estimates. Inspection of the portion of Table 5 dealing with the intercorrelations among standard deviations of estimates reveals that there is a slight relationship between the amount of variability of estimates on all of the factors except DHV. This, it will be noted, is just the opposite of the relationships found for mean algebraic error. On this basis, it appears that those factor not related on directionalized error are related on amount of variability. However, the degree of relationship shown for both of these measures of error are moderately weak, so neither have any appreciable predictive value.

The overall picture of the relationship between direction, magnitude, and variability of errors on the four factors was not consistent, nor was the relationship between error on any two factors consistent over the three measures of error. As a result, although some generality between factors was indicated, particularly for amount of error, when reference is made to the nature of the inspectors estimates, specific reference must be made to a particular factor or factors, and to a specific measure or measures of these estimates.

Intra-Inspector Reliability

In evaluating the reliability of the inspectors estimates of the different factors, considerable information can be obtained by evaluating not only the degree to which these inspectors can agree among themselves on their estimates of the different factors, but also how well they can agree with themselves on their estimates of the different factors when given the opportunity to re-evaluate equivalent sets of samples. To obtain a measure of this agreement,

all of the available inspectors were given a second set of samples to inspect in August. It is the estimates of the 21 inspectors who did evaluate two sets of samples on which the following analyses are based.

Consistency of Estimates

To obtain a measure of the consistency of the relationship between the initial and repeat estimates each inspector gave for the different values of each factor, each repeat inspector's June estimates for each factor were correlated with his August estimates for the same factor. The resultant product-moment correlation coefficient indicates the degree to which each inspector was able to repeat his initial estimates of the different values for each factor. It must be noted that this measure gives no indication of the accuracy or variability of estimates, only the degree to which each inspector tends to estimate repeated equivalent values of each factor in the same way in relation to his estimates of the other values of the same factor.

The results of this analysis are given in Table 6. Probably the most interesting thing about these results is the high degree of consistency all of these inspectors showed in estimating DHV. The degree to which the inspectors were able to repeat their initial estimates of the other three factors was not generally so strong, and disclosed moderate to marked individual differences in this ability for these factors.

Accuracy and Variability of Repeat Estimates

A more critical analysis of the degree of agreement the inspectors achieved with themselves is obtained if the distribution of differences between the June and August estimates is compiled over inspectors for each value of the different factors. These results are presented in tabular form in Table 7. This table presents the mean absolute difference (in per cent) between the June estimates of the 21 inspectors and the August estimates of the same inspectors for each

Table 6

Distribution of Product-Moment Correlation Coefficients Between Initial and Repeated Estimates by 21 Inspectors (N = 16 Samples)

Value of Coefficient	DHV	WOOC	FM	Damage
.95 - .99	17	7	10	2
.90 - .94	4	3	4	4
.85 - .89		2	0	4
.80 - .84		2	6	2
.75 - .79		1	1	3
.70 - .74		2		4
.65 - .69		0		1
.60 - .64		1		0
.55 - .59		1		1
.50 - .54		0		
.45 - .49		1		
.40 - .44		1		

value of the four factors. Also presented are the standard deviation and range of these differences. These data show that these inspectors were unable to duplicate their June estimates when given the opportunity to do so in August. It will be noted that a number of the standard deviations are larger than the mean amount of difference with which they are associated. This is a result of the lack of normality of some of the distributions of differences. The distributions tended to be negatively skewed with a few differences of considerable magnitude, as is shown partially by the different ranges. The considerable magnitude of many of the ranges complements the finding in the previous section which showed that for all factors except DHV there were moderate to marked individual differences in the ability to consistently repeat initial estimates of the different values of these factors. The magnitude and variability of the

Table 7

Mean, Standard Deviation and Range of Differences Between Initial and Repeated Estimates

(N = 21 inspectors)

DHV				Wooc				FM				Damage			
%	Mean	S.D.	Range	%	Mean	S.D.	Range	%	Mean	S.D.	Range	%	Mean	S.D.	Range
12	8.61	6.87	0 - 22.0	0	0.74	1.26	0 - 4.0	0	0.00	0	0 - 0.2	0	0.40	0.68	0 - 2.0
16	9.51	7.55	0 - 26.2	0	0.43	1.15	0 - 5.0	0	0	0	0 - 0	1	0.58	0.65	0 - 2.2
19	5.72	6.46	0 - 24.0	1	1.16	1.29	0 - 4.5	0.5	0.15	0.14	0 - 0.4	2	1.34	1.14	0 - 3.8
23	6.82	6.08	0 - 25.0	2	1.71	1.86	0 - 7.0	0.5	0.20	0.20	0 - 0.7	2	1.15	1.08	0 - 4.4
28	9.82	8.25	0 - 35.2	3	0.84	0.93	0 - 3.0	1	0.23	0.30	0 - 1.2	3	1.20	1.90	0 - 8.8
33	7.62	6.52	0 - 24.0	3	1.79	1.40	0 - 4.4	1	0.40	0.71	0 - 3.1	4	1.39	1.04	0.2 - 5.0
37	6.97	8.41	1.0 - 38.0	4	1.93	1.66	0 - 5.0	1.5	0.33	0.39	0 - 1.4	4	1.30	1.41	0.2 - 5.6
42	9.63	8.52	0.3 - 38.0	5	1.86	2.13	0.2 - 7.0	2	0.32	0.33	0 - 1.5	5	1.61	1.52	0 - 5.7
51	8.44	7.25	0.4 - 26.0	6	1.93	2.73	0 - 12.0	2	0.37	0.46	0 - 1.8	6	1.63	1.32	0 - 5.1
57	6.81	4.60	0 - 19.0	7	1.66	1.72	0 - 6.0	2.5	0.55	0.57	0.1-2.3	7	1.23	1.05	0 - 4.1
64	5.55	3.80	0 - 13.0	8	1.66	1.47	0 - 5.0	3	0.51	0.59	0 - 2.4	8	1.79	1.80	0 - 7.6
74	6.64	7.00	0 - 22.0	8	2.09	2.47	0 - 9.6	3	0.57	1.14	0 - 5.0	10	2.05	1.89	0.8 - 8.2
78	3.64	2.62	0.1 - 9.0	9	2.22	2.52	0 - 10.0	4	0.53	0.60	0 - 4.0	12	4.22	2.86	1.0 - 10.1
83	3.02	3.75	0 - 12.0	10	3.47	3.06	0 - 14.0	5	0.55	0.47	0 - 1.4	13	2.73	2.66	0 - 10.2
89	4.09	3.46	0 - 11.0	10	3.15	2.38	0.2 - 10.6	5	0.83	1.01	0 - 4.5	15	2.99	2.57	0 - 8.5
94	2.38	2.81	0 - 10.0	11	1.85	2.80	0 - 11.4	6	0.52	0.40	0 - 1.2	16	3.08	3.01	0 - 10.2

differences, particularly for DHV, indicate that these inspectors have some rather marked constant and/or variable errors in their estimates of this factor.

To determine the nature and extent of any constant errors, the data were re-analyzed to determine the directionality of the differences found. The results of this analysis are presented in Table 8. In this table, a mean of negative sign indicates thatthe inspectors estimates tended to be lower in August than in June, a mean of positive sign indicates the opposite. These results clarify the situation somewhat, especially for DHV. These inspectors estimates of DHV tended to be somewhat lower in August than in June, but this tendency decreased as the percentage of DHV increased. There was also a tendency for the differences between the inspectors estimates to be somewhat variable, both in amount and in direction as indicated by the standard deviation and range of the different values of DHV. The variability of the differences also tended to decrease as percentage of DHV increased.

The results for each of the other three factors indicate that both the magnitude and variability of the differences in estimates tended to increase as the programmed values of these factors increased, but that there was no consistent tendency for these differences to be either higher or lower in August than in June.

Changes and Repeated Errors in Grade and Sub-Class

The number of changes, both single and multiple, and repeated errors in grade and subclass resulting from these inspectors repeated estimates of equivalent values of the different factors are presented in Table 9. The percentage values at the bottom of the different columns represent the number of occurrances of changes or repeated errors divided by the total number of

Table 8

Algebraic Mean, Standard Deviation, and Range of Differences Between Initial and Repeated Estimates. (N = 21 Inspectors)

DHV				WOOC				FM				Damage			
%	Mean	S.D.	Range	%	Mean	S.D.	Range	%	Mean	S.D.	Range	%	Mean	S.D.	Range
12	-4.94[1]	10.00	-22.0 - +18.0	0	-0.17	1.47	-3.0 - +4.0	0	0	0	0	0	0.25	0.76	-1.5 - +2.0
16	-6.89	10.27	-26.2 - +10.0	0	0.21	1.22	-2.0 - +5.0	0	0	0	0	1	-0.03	0.93	-1.5 - +2.2
19	-5.57	6.59	-24.0 - +1.0	1	-0.30	1.73	-4.5 - +2.9	0.5	-0.06	0.17	-0.4 - +0.4	2	0.12	1.78	-3.7 - +2.0
23	-5.42	7.44	-25.0 - +10.0	2	-0.23	2.56	-7.0 - +4.5	0.5	-0.10	0.26	-0.7 - +0.4	2	-0.36	1.56	-4.4 - +2.2
28	-9.06	9.14	-35.2 - +5.6	3	0.18	1.29	-3.0 - +3.0	1	0.02	0.39	-1.2 - +0.8	3	0.29	2.24	-2.1 - +8.8
33	-5.25	8.64	-24.0 - +13.0	3	0.63	2.22	-3.0 - +4.4	1	-0.26	0.77	-3.1 - +0.8	4	-0.17	1.75	-5.0 - +2.1
37	-5.93	9.20	-38.0 - +3.0	4	1.00	2.40	-4.5 - +5.0	1.5	-0.02	0.52	-1.2 - +1.4	4	-0.25	2.11	-4.4 - +5.6
42	-5.37	11.97	-38.0 - +12.0	5	-0.27	2.85	-7.0 - +7.0	2	0.00	0.47	-0.7 - +1.5	5	-0.95	2.13	-5.7 - +2.2
51	-4.09	10.50	-26.0 - +25.0	6	-0.75	3.29	-12.0 - +4.4	2	0.22	0.55	-0.6 - +1.8	6	-0.41	2.09	-5.1 - +3.0
57	-5.38	6.22	-19.0 - +10.0	7	0.36	2.36	-3.0 - +6.0	2.5	0.34	0.73	-0.6 - +2.3	7	-0.94	1.33	-4.1 - +1.6
64	-2.55	6.35	-13.0 - +8.0	8	0.30	2.23	-5.0 - +4.6	3	0.06	0.75	-1.7 - +2.4	8	-0.01	2.57	-4.9 - +7.6
74	-1.33	9.67	-22.0 - +20.0	8	0.66	3.20	-9.6 - +6.0	3	0.00	3.20	-5.0 - +2.5	10	0.01	3.56	-8.2 - +4.8
78	-1.53	4.39	-9.0 - +6.0	9	0.02	3.38	-6.5 - +10.0	4	-0.16	0.81	-4.0 - +1.9	12	2.38	4.43	-5.6 - +10.1
83	0.00	4.87	-12.0 - +10.4	10	-0.56	4.65	-14.0 - +6.4	5	-0.14	1.32	-1.7 - +4.5	13	-0.95	3.74	-10.2 - +5.7
89	-1.11	5.32	-11.0 - +10.0	10	-0.36	4.00	-5.0 - +10.6	5	-0.03	0.72	-1.4 - +1.4	15	-0.21	3.97	-8.5 - +6.8
94	-0.34	3.70	-8.0 - +10.0	11	-0.40	3.35	-11.4 - +6.5	6	0.26	0.61	-1.0 - +1.2	16	0.15	4.36	-9.1 - +10.2

[1]Sign of mean indicates direction of mean difference, + indicates that August estimate higher, - indicates that August estimate lower than June estimate.

determinations of that factor. The estimates for the three grade factors were first evaluated separately, as though each factor was the grade determining factor in every case, then the results of the composite grade resulting from the inspectors estimates for the three grading factors combined. It should be remembered thatthe grade and subclass of a sample are completely independent of each other, so no combined factors analysis of the subclass estimates was required.

Table 10 presents the number of changes and repeated errors which occurred in estimation of grade and/or subclass of the entire sample. The results in this table are of the type kept by the Federal Grain Supervisors for inspections of commercial samples at two or more inspection points, and for samples inspected by both licensed grain inspectors and Federal Grain Supervisors.

The results presented in Tables 9 and 10 indicate that these inspectors repeat estimates did not always place each of the factors in the same grade or subclass as their initial estimates did, and that at least some of the time neither estimate of a particular value of a factor placed that factor in the proper grade or subclass, (as shown by the column listed repeated errors). The percentage figures given at the bottom of the columns cannot be added to obtain total number of errors. In the tabulation of the results, if an inspectors two estimates of any particular value for any factor placed that factor in the wrong grade or subclass both times, but in a different wrong grade or subclass each time, then the inspector was considered to have made both a repeated error, and also a change or multiple change in grade or subclass. As a result, the percentages of changes and multiple changes in grade or subclass for any factor may be added to obtain the total number of changes, but the percentage of repeated errors cannot also be added.

The results for subclass (when considered independently) indicates that

Table 9

Summary of Changes and Repeated Errors in Grade and Subclass for Replicated Estimates. (N = 21 inspectors)

Master Sample No.	Changes and Repeated Error for Subclass: Per Cent	Chg	Multiple Chgs	Repeat Errors	N	Changes and Repeated Errors in Grade for Each Grade Factor: WOOC Per Cent	Chgs	Multiple Chg	Repeat Errors	N	FM Per Cent	Chgs	Multiple Chg	Repeat Errors	N	Damage Per Cent	Chgs	Multiple Chgs	Repeat Errors	N	Changes and Repeated Errors in Grade of Composite Sample: Chgs	Multiple Chgs	Repeat Errors	N	Grade[b]	Determining Factor & Value
1	64	8	0	0	20	8	4	0	0	21	6	7	0	3	21	15[a]	7	3	2	21	7	0	2	21	6	FM (6%)
2	16	4	0	0	21	6	5	0	2	21	1[a]	4	0	0	21	13	8	4	6	21	7	6	6	21	5	Damage (13%)
3	19	0	0	0	19	9	10	0	3	21	2.5	6	3	2	21	2[a]	5	0	0	21	2	10	6	21	4	FM (2.5%)
4	94	0	0	0	21	0	0	0	0	21	0	0	0	0	21	3	7	1	2	21	7	1	2	21	2	Damage (3%)
5	89	0	0	0	20	0[a]	0	0	0	21	5[a]	1	0	0	20	1[a]	1	0	1	20	1	0	0	20	5 or 6[c]	FM (5%)
6	78	1	0	0	20	5[a]	1	0	0	20	5[a]	2	0	0	20	7[a]	2	0	0	20	2	0	0	20	5 or 6	FM (5%)
7	23	3	0	0	20	4	8	0	1	20	0.5[a]	0	0	0	12	4[a]	7	0	1	20	3	5	1	20	2 or 3	Damage (4%)
8	57	0	0	0	21	3	6	0	0	21	3[a]	6	1	1	21	4[a]	4	1	0	21	7	0	1	21	4 or 5	FM (3%)
9	51	2	0	0	20	7	3	1	0	20	4	5	0	0	20	5	5	3	1	20	4	0	1	20	5	FM (4%)
10	28	7	0	2	20	8	2	0	4	21	3[a]	4	1	0	21	16	8	2	16	21	10	0	12	21	6	Damage (16%)
11	33	8	0	3	21	2	5	0	0	21	2[a]	1	1	0	21	10[a]	11	2	6	21	10	0	6	21	4 or 5	Damage (10%)
12	74	7	0	13	21	3	1	0	0	21	2[a]	1	1	0	21	8	10	1	14	21	6	1	14	21	4	Damage (8%)
13	42	3	0	0	21	1	0	0	0	21	1[a]	5	0	0	20	2[a]	1	0	0	21	2	1	0	21	2 or 3	FM (1%)
14	37	8	0	8	19	10[a]	3	0	0	21	1.5	6	1	1	21	12	9	8	11	20	3	10	3	20	5 or 6	WOOC (10%) Damage (12%)
15	83	1	0	0	17	11[a]	7	1	0	21	0.5[a]	0	0	0	21	0	0	0	0	21	0	7	0	21	6	WOOC (11%)
16	12	0	0	0	18	10[a]	1	0	0	21	0.0	0	0	0	21	6	12	1	4	21	1	2	0	21	3 or 6	WOOC (10%)
Σ		52	0	26	319		56	2	10	333		48	8	7	323		97	26	64	331	72	43	54	331		
%		16.4	0	8.2			16.8	0.6	3.0			14.8	2.4	2.1			29.3	7.9	19.3		21.7	13.0	16.3			

[a]Values which fall on grade cutting points.

[b]Grade for sample is lowest grade (numerically highest) received by any grading factor.

[c]When value of grade-determining factor falls on a grade cutting point, grade for that sample was considered correct if estimate fell in normally proper grade or next grade above it.

Table 10

Summary of Changes and Repeated Errors in Grade and/or Sub-Class in Repeated Estimates of Composite Sample (N = 21 Inspectors)

Master Sample No.	Changes in Grade and/or Sub-Class	Multiple Changes in Grade and/or Sub-Class	Repeated Errors in Grade and/or Sub-Class	N	Grade[a]	Grade Factor & Value	Sub-Class	DHV
1	13	0	2	20	6	FM (6%)	Hard Winter	(64%)
2	7	6	7	21	5	Damage (13%)	Yellow Hard	(16%)
3	2	8	6	19	4	FM (2.5%)	Yellow Hard	(19%)
4	7	1	2	21	2	Damage (3%)	Dark Hard	(94%)
5	1	0	0	20	5 or 6[b]	FM (5%)	Dark Hard	(89%)
6	3	0	0	20	5 or 6	FM (5%)	Dark Hard	(78%)
7	3	5	1	20	2 or 3	Damage (4%)	Yellow Hard	(23%)
8	7	0	1	21	4 or 5	FM (3%)	Hard Winter	(57%)
9	5	0	1	20	5	FM (4%)	Hard Winter	(51%)
10	14	0	13	20	6	Damage (16%)	Yellow Hard	(28%)
11	9	4	7	21	4 or 5	Damage (10%)	Yellow Hard	(33%)
12	11	0	17	21	4	Damage (8%)	Hard Winter	(74%)
13	5	1	0	21	2 or 3	FM (1%)	Hard Winter	(42%)
14	6	7	8	18	5 or 6	WOOC (10%) & Damage (12%)	Yellow Hard	(37%)
15	1	7	0	17	6	WOOC (11%)	Dark Hard	(83%)
16	2	0	0	17	3 or 6	WOOC (10%)	Yellow Hard	(12%)
Σ	96	39	65	317				
%	30.2	12.3	20.5					

[a]Grade of sample is lowest grade (numerically highest) received by any grading factor.

[b]When value of grade determining factor falls on grade cutting point, grade of that sample is considered correct if estimate fell in either the normally correct grade or the next grade above it.

most of the changes and repeated errors in the estimation of sub-class occurred for values of DHV just below the subclass cutting points. Five of the 16 values of DHV (28 per cent, 33 per cent, 37 per cent, 64 per cent, 74 per cent) accounted for 38 of the 52 changes in subclass. Four of these values, (28 per cent, 33 per cent, 37 per cent, and 74 per cent DHV) account for all 26 of the repeated errors of subclass. These results, in conjunction with the tendency of these inspectors August estimates of DHV to be lower than their June estimates as shown in Table 8 indicate that in August some of these inspectors continued to overestimate the value of DHV to a sufficient degree to maintain the estimate of the subclass of the four values mentioned one subclass too high. They also indicate that for other inspectors the amount of overestimation of those four values was lowered sufficiently to change the estimated subclass.

Damage is the only factor that shows any sizable percentage of repeated errors, and inspection of Table 9 will show that the majority of these repeated errors occurred on large values of damage or values of damage just above a grade cutting point. This indicates that these inspectors were, to a fair degree, consistently underestimating the amount of damage both in June and August.

The 34.7 per cent total changes of grade alone, and the 42.5 per cent total changes in grade and/or subclass clearly indicate that these inspectors were not able to consistently evaluate the samples in the same grade and subclass on repeated evaluations of equivalent samples.

Situational and Personal Data Correlates of Accuracy and Reliability

Factors such as amount of light, experience, age, general working conditions, etc. have been shown by numerous studies to be important factors contributing to the accuracy and consistency with which visual discriminations can

be made. With this in mind, data were collected on age and length of service of the inspectors, and light meter readings were taken at the time each sample was inspected. This section deals with the relationship between these factors and the accuracy and consistency of the inspectors determinations.

Age

It would appear that age might have an effect upon the nature of the estimates made by an inspector, in that with increasing age fatigue would be much harder to combat, and particularily important would be any loss of visual acuity. To evaluate any effect age might have, three product-moment correlations were run between age and three different measures of the inspectors estimates of each factor. Thesestimates of the initial set of samples inspected by 36 of the 40 inspectors for whom age data were available were used in this analysis, and all other analyses in this section. The first correlation was between age of the inspector, and each inspectors mean amount of error for each factor based on his 16 estimates of that factor. This would give the relationship between age and the average amount of error, per se, that an inspector tends to make. The second correlation was run between age and the mean algebraic error of the 16 estimates of each factor by each inspector. This correlation indicates the relationship between age of the inspector and the direction in which errors tended to be made. A third correlation was run between age of the inspector, and the standard deviation of the errors on each factor. This correlation would give the relationship between age and variability of the inspectors estimates of each factor.

The results of these correlations are presented in Table 11. It will be noted that there is no consistent relationship between age and any of the three measures of error for any factor, nor does age appear to be consistently related to all factors on any of the three measures of error. Mean amount of error

and mean algebraic error on DHV and damage show a moderate degree of relationship to age, but the variability of neither appears to bear any relationship. There appears to be no appreciable degree of relationship between age and WOOC or FM on any of the measures of error. The variability of error on all factors appears to be least related to age.

Table 11

Product-Moment Correlation Coefficients Between Age of Inspector and Each of Three Measures of Error for Each Factor (N = 36 inspectors)

	Mean Amount of Error	Mean Algebraic Error	Standard Deviation of Error
Age x DHV	.37*	.36*	.08
Age x WOOC	-.11	-.08	.15
Age x FM	.26	-.15	.30
Age x Damage	.38*	-.30	.20

*Significant at .05 level ($r_{.05} = .329$)

Length of Service

It also appeared that amount of experience as an inspector might be related to the nature of the estimates given for each factor. To evaluate this possibility, data were collected on the length of service as an inspector (in years), for 39 of the 40 inspectors.

The same procedure was followed here as was used for age. The results of the correlations are shown in Table 12. Generally, there is even less relationship between length of service and the three measures of error than was found for age. Again, mean algebraic and mean amount of error on DHV and mean algebraic error on damage appear to have some relationship. It is interesting to note that as experience increases, the inspectors tend to make a larger amount of error on DHV, to overestimate the value of DHV more, and to underestimate

the value of damage more. (A similar relationship was found for age.)

Table 12

Product-Moment Correlation Coefficients Between Length of Service and Each of Three Measures of Error for Each Factor (N = 39 Inspectors)

	Mean Amount of Error	Mean Algebraic Error	Standard Deviation of Error
Length of Service x DHV	.31	.34*	.06
Length of Service x WOOC	-.22	-.23	-.09
Length of Service x FM	.12	-.12	.12
Length of Service x Damage	.19	-.36*	.04

*Significant at .05 level ($r_{.05}$ = .320)

Amount of Light

Proper evaluation of the factors investigated require fine discriminations based on visual cues. It would thus appear that the amount of light under which these inspections are made would be of considerable importance. All of the inspectors who cooperated in this study evaluated these samples under "natural" light, usually from a north window. Amount of light available as recorded by light meter readings varied considerably (range = 2.8 to 32.0, mean = 12.6).

To evaluate the effect, if any, amount of light had on the nature of the estimates made by the inspectors, product-moment correlations were run between the meter reading for each sample and the estimate the inspector gave for one of the factors in that sample. All of the inspectors first estimates of one value of that factor constituted the N for the correlation (all inspectors estimates of one factor of one master sample). One value for each factor was randomly picked from the low, low-medium, high-medium, and high quartiles of the

range for that factor. This resulted in 16 correlation coefficients, four for each factor. The results of this procedure are shown in Table 13. There appears to be no appreciable or consistent relationship between amount of light and the estimates of any of the values of the four factors. The distribution of correlation coefficients appears to be random and can be readily attributed to chance.

Table 13

Product-Moment Correlation Coefficients Between Light Meter Reading and Inspectors Estimates of Each of Four Selected Values of Each Factor
(N = 30 to 34 Inspectors)

	Quantile of Range in Which Selected Sample Fell			
	Low	Low Med	High Med	High
Meter Reading x DHV	-.28	.03	-.07	.23
Meter Reading x WOOC	-.06	.05	-.18	.14
Meter Reading x FM	.21	.52**	-.11	-.09
Meter Reading x Damage	-.29	-.04	-.15	.34

**Significant at .01 level of significance ($r_{.01} = .409$), ($r_{.05} = .349$)

Station Differences

Another environmental factor which might influence the estimates made by the inspector is the general physical and social surroundings and working conditions under which the inspector is working. To evaluate any differences due to station a between-within analysis of variance was run on the average absolute amount of error made on each factor by the inspectors at each of the seven stations where there were three or more inspectors. The results of this analysis are given in Table 14. A significant F was obtained for all of the factors except WOOC. This indicates that the differences between inspectors in mean amount of error made on all factors except WOOC were greater between stations than they were within individual stations.

Table 14

Summary of Analyses of Variance to Evaluate Difference Between Station on Amount of Error

	Source of Variance	d.f.	Sum of Squares	Mean Square	F
DHV	Between Stations	6	228.64	38.10	7.83**
	Within Stations	30	145.86	4.86	
WOOC	Between Stations	6	4.76	0.79	1.61
	Within Stations	30	14.75	0.49	
FM	Between Stations	6	1.06	0.17	5.66**
	Within Stations	30	1.13	0.03	
Damage	Between Stations	6	9.96	1.66	4.04**
	Within Stations	30	12.42	0.41	

**Significant at .01 level

DISCUSSION

In this section, the results presented in the previous section will be discussed first for each factor separately, and then combined. This approach was taken, rather than discussion of the results in terms of inter- and intra-inspector reliability as was previously done. The primary reason for this is the fact that for any given inspector, the direction, magnitude, and variability of errors made on one factor are not consistently related to the nature, magnitude, and variability of errors made for any other factor involved in the inspection process. This was shown by the inter-correlation of estimates for the four factors investigated. Although some of these inter-correlations are moderately large, there is no consistent relationship of any appreciable size between all factors for any one of the three measures used; there was also no consistent relationship of any magnitude for any one factor on all three measures used. As a result, reference can not be made to any particular

inspector as being "good" or "bad" in respect to the reliability of his estimates generally. Reference must be made to some particular measure of the reliability of estimates for one particular factor.

Sub-Class

The estimates of subclass by the individual inspector in June and again in August showed a fairly high degree of consistency, as was indicated by the high values obtained when each repeat inspectors estimate of the percentage of DHV for June and August were correlated. This relationship does not necessarily indicate identity of estimates, but rather may indicate consistency of proportionality of estimates. The latter is suggested by the mean algebraic differences between the June and August estimates of the 21 repeat inspectors for DHV which show that the individual inspector tended to lower his estimates of DHV in August to a greater extent, the lower the value of DHV. However, the initial estimates of the different values of DHV, and the amount of the difference between the June and August estimates of DHV were not of equal magnitude for the different inspectors. This is indicated by the standard deviation and range of the estimates of each value of DHV in June, and the standard deviation and range of the differences in repeated estimates of DHV.

It may also be noted that as the age and length of service of the inspector increased, there was a moderate tendency for both the mean algebraic and mean absolute error of the estimates of DHV to be higher. On the other hand, the variability of the estimates of DHV by the inspector was not appreciably related to either the age or length of service of that inspector. One interpretation of this tendency is that the inspector, with continued association with grain merchandisers, becomes progressively more lenient in his evaluation of these

factors, thus to a greater extent giving the consigner the benefit of the doubt. The amount of error in estimating per cent DHV was related to the station at which the inspector was located, but the estimates of DHV showed no appreciable relation to the amount of light available. The lack of relationship to light indicates that within the range of light intensities found in this investigation, these inspectors were able to maintain perceptual constancy, at least to a limited extent. The present data do not allow a definitative evaluation of the pehnomenon. Nor do they permit any evaluation of other criteria of efficiency of performance in relation to illumination, that is eyestrain, time to complete inspections, etc.

Both the variability in the initial estimates, and the changes in the estimates of DHV from June to August resulted in a fairly sizable percentage of the samples inspected being placed in an improper subclass on one or both occasions. (Initial errors 19 per cent, changes 16 per cent, and repeated errors 8 per cent.)

The lowering of the mean estimate of DHV from June to August is unexplained, and rather mystifing. One possible explanation is that the nature of the large number of samples inspected by the inspectors in the intervening time period resulted in some pronounced adaptation effects which influenced their estimate of DHV.

Grading Factors

Wheats of Other Classes

The individual inspectors showed the least amount of uniformity in the ability to make consistent estimates of this factor, as shown by the range of values obtained when the June and August estimates of WOOC for each repeat

inspector were correlated. The inspector's initial mean estimate, the variability and range of estimates of WOOC, and the direction of the differences of repeated estimates of this factor showed no consistent trend over the programmed values of the factor. The mean amount of the difference between repeated estimates of this factor did tend to increase as the programmed value increased. Part of the lack of consistency in direction and magnitude of mean error over the range of values programmed was accounted for by the fact that these inspectors tended to confuse Yellow Berry kernels with kernels of WOOC, and this tendency increased with the amount of YB kernels in the sample ($r_{12.3} = -.65$).[4] Otherwise, the amount and direction of mean error, and the variability of estimates showed no appreciable relationship to age, length of service, amount of light available, or station at which the samples were inspected.

The initial estimates and differences in estimates between June and August resulted in a relatively small per cent of the samples inspected being placed in improper grades (14 per cent initial errors, 16 per cent changes, and 8 per cent repeated errors). The comparatively large range for each grade (5 per cent in each case) and the fact that there were three programmed values which fell on grade cutting points (5 per cent and 10 per cent), which resulted in larger possible errors in estimates without changes or errors in grade, can be offered in partial explanation of this finding.

The fact that estimates of this factor depended upon the amount of YB kernels in the sample, and did not show any orderly relationship to any of the other factors investigated, or to age, length of service, or station differences indicated that the conditions controlling the estimation of values of this

---

[4]The partial correlation was obtained between mean estimate of WOOC and programmed value of DHV, with per cent WOOC partialed out, but per cent DHV and per cent YB have a perfect negative relationship, so the correlation may be interpreted either in terms of per cent DHV or per cent YB.

factor are unique among the factors investigated. What these conditions are, and why these conditions are different than those found for the other factors is not ascertainable from the data available at the present time. Any discussion of these conditions would be purely speculation, and there is not sufficient information to merit even that.

Foreign Material

The consistency with which estimates of this factor were repeated by individual inspectors, and the uniformity of this ability over the 21 repeat inspectors was second only to the consistency and uniformity shown for DHV. The increasing variability, the progressively greater mean underestimation of the programmed value, and the magnitude of the differences in repeated estimates as amount of FM increased, indicate that the initial and repeat estimates were not equally accurate for the different inspectors, or for the different values of FM for any one inspector. Furthermore, neither direction, magnitude, nor variability of these estimates was appreciably related to the age or length of service of the inspector, or the amount of light under which the estimates were made. The station at which the inspector was located did make a difference in the mean amount of error made.

The moderate percentages of initial errors (12 per cent), changes (17 per cent), and repeated errors (2 per cent) in grade made by these inspectors for this factor are considerably smaller than would be expected from viewing the distribution of initial estimates and difference in repeated estimates. Again, this can be partially explained by the fact that 10 of the 16 values programmed fell on grade cutting points which resulted in a much more conservative estimate of percentage of samples misgraded.

Damage

The correlations computed as indices of the individual inspector's

consistency in repeated estimates of this factor showed that these inspectors were less consistent in their estimates of the different values of damage than for any other factor investigated. Although the range of correlation values is not as large as that for WOOC (Table 6) the majority of these values are lower than those for WOOC. The considerable amount of variability among the initial estimates of Damage and the fairly consistent and progressively greater mean underestimation of Damage as the amount of damage increases, as well as the differences in repeated estimates, taken together indicate the lack of consistency of the inspectors in making estimate of damage. The partial correlation computed to indicate the relationship of per cent sick or black germ damage to the error of estimation of total damage ($r_{12 \cdot 3} = -.88$) shows that this type of damage presented the primary difficulty for the inspectors in evaluating total damage. There was a moderate tendency for the amount of error on damage and amount of underestimation of damage to increase as age and length of service of the inspector increased. The station at which the inspection was made was associated with the amount of error, but amount of light had no appreciable or consistent effect.

The relationship between station of the inspector and the mean amount of error made on all factors except WOOC could be due to a number of things, among them are quantity and quality of light and color of background or working surface, and of course the differences in ability of the inspectors at the different stations. Anotherpossibility is that the differences between stations reflect the influence of the varying social psychological environments of the different stations. Reference is made here to such things as attitudes toward work, employer-employee relationships, local norms, both in reference to the standards and to the need for accuracy, and general morale.

The fairly marked lack of consistency and accuracy in the estimation of

the amount of damage in the samples is further brought out by the percentages of errors in the initial estimates (31 per cent), and of changes (37 per cent), and repeated errors (19 per cent) from June to August in grade when damage is considered alone. These percentages are conservative estimates of the inspector's reliability in relation to grade, since seven of the 16 values of damage fell at cutting points between two grades and either grade was accepted as correct.

One possible explanation of the difficulty in the estimation of damage is the fact that the grain standards appear to impose a dichotomous classification on what is essentially a continuous variable, especially in respect to sick or black germ damage. According to the definition given by the Official Grain Inspection Manual (United States Department of Agriculture 1952) "Kernels damaged by heat --- Kernels which are damaged as a result of heat but which are not materially discolored, shall be damaged kernels." Sick wheat comes under this classification and refers to kernels whose germ has been discolored or turned brown due to heating. The browning of the germ is a continuous process resulting in continuously varying shades of darkness. The difference between sick and non-sick wheat appears to be a matter of degree of browning of the germ, and not an all-or-none type of situation. In a situation such as this, it is not at all surprising thatindividual inspectors are not able to maintain consistent estimates of this factor or that different inspectors are not able to agree on what is and what is not sick wheat.

Composite Grade of Sample

As has already been presented, the grade of the sample is determined by the factor or factors which fall in the lowest grade. The percentage of samples which were changed in grade (35 per cent) or were incorrectly graded both times (16 per cent) or were graded improperly the first time (31 per cent) have been

represented as a conservative estimate of the reliability of licensed inspectors in the estimation of grade. However, it can be argued that just the opposite is true, that is, that these percentages represent overestimation of the percentage of samples misgraded by inspectors in their every day work. This argument is based on two points: First, in the inspection of most commercial samples, the inspector is working with samples that have only one, or at most two factors that might lower the grade. In the experimental sample sets, most of the samples contained percentages of all three grade-determining factors. This in turn increased the probability of an error in grade occuring, particularly on samples such as 1 and 14, which contain high percentages of two grading factors. The second point is that, according to the records kept by the Agricultural Marketing Service of the USDA, about two-thirds of the samples of Hard Red Winter Wheat inspected in the United States each year are grade No. 1, while only about three per cent of the samples inspected each year are placed in grade 4 or lower (United States Department of Agriculture 1959). Yet 13 of the 16 samples used in this study fell in the lowest three grades with none falling in grade No. 1. This then, presented an unfair test of the inspectors ability to estimate grade properly, because they do not evaluate such a large proportion of complex samples of low grade in their everyday work.

These factors do not invalidate the results of the study for two reasons. First, no one knows exactly how much of each factor is in any commercial sample. If a carload of grain was prepared as each of these samples was prepared, and the inspectors evaluated a sample from it, the values of the different factors would still not be known because only a sample of the carload is evaluated by the inspector, and any differences between inspectors estimate of the sample, and the values programmed could be due to sampling error. For this source of error to be eliminated, the inspector would have to inspect the entire carload

of grain. Because of this fact, the degree to which the commercial samples actually do fall in the higher grades, and the degree to which the high percentage of samples placed in the upper grades reflect oversights or underestimations of grade determining factors can not be determined. Second, none of the samples programmed presented a situation that could not naturally occur, and therefore a situation that a licensed grain inspector would not face at some time with a commercial sample.

The interpretation given these results, then, is that for these samples, the percentage figures obtained for samples misgraded are conservative estimates because of the allowance of two correct grades for those samples in which the value of the grade-determining factor fell on a grade cutting point. However, caution should be used in generalizing from the summary percentages obtained, to statements about the day-to-day accuracy of the inspectors because of the high proportion of samples of lower grades in the sample sets.

Composite Grade and Sub-Class of Sample

Here again, the percentages of initial errors (40 per cent), changes (43 per cent), and repeated errors (21 per cent) in grade and/or subclass of the samples as a whole do not indicate a high degree of reliability for these inspectors, at least on these samples. It will be noted that the percentages here are higher than for either subclass or composite of grade alone, but are lower than the two summed together. This of course is to be expected, because grade and subclass are independent, and also, because an error on both grade and subclass on the same sample was counted as one error only.

There are at least three possible explanations that can be offered to account for the lack of accuracy and consistency found in this study. The

first explanation was presented previously in the section on grade of the composite sample. The explanation was essentially that the lack of reliability of the inspectors was due to the experimental samples being too complex, and mostly in the lower grades. This constituted an unfair test of their ability.

The second explanation is that different inspectors, as the result of different training, personal interpretation of the standards and different local or station norms are using different standards to evaluate the samples, and the unreliability found was due to these individual differences in interpretation of the standards. This explanation has merit in that the high degree of consistency in repeated estimates of DHV and FM support this view, as do the initial errors and repeated errors of the different factors in grade or subclass. The significant differences between stations on amount of error made on DHV, FM, and Damage, if interpretated as indicating differences between individuals only, also support this view. However, this explanation cannot adequately account for the lack of consistency of repeated estimates for WOOC and Damage, or the differences in repeated estimates and the resultant changes in grade or subclass.

A third possible explanation is that the standards are not defined in such a way that the inspectors can consistently and accurately evaluate the different factors in the manner required. The basic assumption here is that the inspectors are not capable of responding consistently and accurately to the limited differences in brightness, hue or morphological characteristics required by the standards, and the imposition of arbitrary cutting points on essentially continuous gradations. Examples were indicated in the determination of WOOC and Sick Damage. The definitions given in the Grain Inspection Manual are not highly specific in regard to characteristics whereby the various determinations are to be made. An example was given for sick or black germ damage.

This explanation can adequately account for all of the data obtained except

the high degree of consistency of repeated estimates shown for DHV and FM, and the progressive increase or decrease in the differences between initial and repeated estimates in mean over-or underestimation of DHV, FM, and Damage with increasing values of these factors.

All three explanations offered have merit and the last two can account for a considerable proportion of the results obtained. On the basis of the data now available, none of the explanations can be disproven, so evaluation of the relative merits of the different views will have to await further investigation.

Evaluation of the Study

The present study was designed to obtain a quantitative description of the inter- and intra-inspector reliability and accuracy of four subjective judgments required of practicing licensed grain inspectors in the inspection of Hard Red Winter Wheat when working under existing field conditions.

The technique used to prepare the samples for this study resulted in relatively small error. The values of the four factors investigated were undoubtedly as accurate as could be achieved with this technique or any other technique yet devised. The one question which remains unanswered is the degree to which the different samples were equivalent in the quality of difficulty of each factor programmed. The population of each factor was as homogeneous as possible, but the lack of identity of the factors in each sample could have had an effect. Only future research can evaluate any effect this might have had.

The lack of resources, the number of inspectors available, and the fact that the study was done under field conditions restricted the refinement with which some of the factors were evaluated. The lack of a sufficient number of inspectors removed the possibility of using a more powerful experimental design.

The results for the inspectors estimates for any of the four factors have to be evaluated in the context of the entire sample, since the values programmed for the other three factors were not held constant over the range of any one factor.

The evaluation of the effect of amount of light, station, age, and length of service are only secondary analyses and do not necessarily represent an accurate evaluation of these effects, since variations in one of these factors were accompanied by changes in one or more of the other factors evaluated. Further research must be done to evaluate any of the particular factors under constant conditions.

The most serious objection to the study is that the results may have no generality for the evaluation of grain in general. The major basis of this objection is that the nature or characteristics of the four factors investigated, and thus the nature and difficulty of the judgments required in evaluating them, vary considerably from locality to locality, and from season to season. The contention is, that to arrive at any representative evaluation of the reliability of the judgments of the factors investigated requires many and varied characteristics or qualities of the populations of the factors investigated.

It should be noted, on the other hand, that the particular values of the mean under- or overestimation of any particular value of any factor would be a function of the particular population used in the study. However, it is very difficult to see how the obtained systematic changes in the variability and mean of the inspectors estimates over the range of any factor can be attributed to the characteristics of the particular population used for any of the factors investigated.

Two other comments should be made about the study, the first is that there is the possibility that some of the errors in estimation are due to weighing and calculating errors made by the inspector, and are not errors in judgment. For

example, some estimations were noted which were almost exactly twice or half of the correct value, suggesting that here an error in arithmetic was made. In other words, the procedures used in this study did not permit the isolation of errors of judgment from those of weighing and arithmetic. The other comment is that the fact that some of the inspectors refused to use the entire sample to evaluate all or some of the factors introduced a source of error due to sampling which cannot be isolated from the errors in judgment made by the inspector. The view taken on both of these conditions is that if the study were to be redone every effort would be made to remove the confounding of these different types of errors. On the other hand, both of these conditions would affect the final evaluations of any routine samples inspected by these inspectors, and do not represent sources of error unique to the experimental samples. Any difference found between the accuracy and reliability of inspectors who inspected the entire sample, and the inspectors who did not, could not be attributed to the procedure of cutting the sample alone, because this is confounded with difference in the ability of the inspectors who were doing the inspecting.

Implications for Future Research

The general finding that these inspectors were not able to agree with themselves or other inspectors on the estimation of the programmed values of the factors investigated, and further that these differences in estimation resulted in fairly sizable percentages of the samples being placed in wrong grades or subclasses indicates that further research is definitely needed. The nature of future research can properly take either of two courses, one, the conclusion can be made that these determinations are too difficult for the

human to make and objective chemical and mechanical techniques might be developed to replace the present determinations. (In this connection, a number of such techniques which might be used have already been developed.) Second, research can be directed toward identifing the factors which are contributing to errors in estimation of the different factors and removing these to the extent possible; in other words, maximizing the conditions under which these determinations are made.

The first type of approach is properly the concern of the cereal chemist and milling technologist; the second approach may properly be the concern of the psychologist.

Further studies, directed toward identifing the factors affecting the inspectors estimates of these subjective factors, should be carried out in a laboratory situation in which as many of the factors that can be controlled, are controlled. Such factors would be age and length of service of the inspector, quantity and quality of light, color of background surface, and distractions and interruptions during the inspection process. All of which are difficult or impossible to control in a field situation.

There are a number of factors which will have to be investigated before a clear picture of what the maximum conditions for these determinations are is achieved. It would appear that an evaluation of the effect of quantity and quality of light and the color of the working surface on which these determinations are made should be one of the first factors investigated. No significant relationship was found between inspectors estimates and amount of light available in the present study, but in this evaluation, effect of amount of light was confounded with any changes in quality of light, differences in color of working surface, inspector making the estimate, and other factors which might well have been confounding factors.

Another factor deserving investigation is adaption effects. The tendency of the inspectors to estimate DHV lower in August than in June might well have been due to just such an effect. The quality of grain inspected in the field situation is not constant, but changes from sample to sample, and also from season to season. It would be of considerable value to know if the degree to which tendencies of the inspectors to over- or underestimate the different factors is due to the nature of the contrast between the samples the inspector is presently inspecting, and the samples the inspector is accustomed to inspecting.

Another factor of considerable importance, is the amount of time available to make the determinations. In the present study no limits were set, as the inspector was instructed to work at his own pace. Due to the fact that most of the inspectors who cooperated in the study were frequently interrupted during the time they were inspecting the experimental samples, no index of the relationship between speed of determination and accuracy and reliability of estimates was obtainable. In the actual field situation, the time available to inspect each sample varies from day to day, and season to season. If the minimum amount of time within which most inspectors could make accurate determinations was determined, it might well reduce the inaccuracy and inconsistency which the inspectors make these determinations.

Other problems which should also be investigated include the effect of the value of the other factors in the sample on the estimate of each of the factors, accuracy and reliability of the determination of odors, relationship of age and length of service of inspector to accuracy and consistency, effect of the quality or type of population used for each factor determined, and possibly some other topics which may become evident as the research progresses.

The present study has raised many more questions than it has answered, it

did what it was designed to do, and that was to evaluate the present status of the accuracy and reliability of the subjective judgments made by licensed grain inspectors under field conditions.

SUMMARY

This report presents a psychophysical investigation of the inter- and intra-inspector reliability and accuracy of four subjective determinations made in the inspection of Hard Red Winter Wheat by practicing licensed grain inspectors under field conditions. The criteria used to evaluate the accuracy and reliability of these determinations was the standards given in the United States Grain Standards Act. Equivalent sample sets of sixteen samples each, containing accurately programmed values of Dark, Hard, and Vitreous (DHV), Wheats of Other Classes (WOOC), Foreign Material (FM), and Damaged (Damage) kernels, were prepared. These sample sets were then administered to a total of 40 practicing licensed grain inspectors at nine different inspection points throughout Kansas and Missouri on two different occasions. On the first occasion 29 inspectors completed sample sets, on the second occasion, approximately two months later, 21 of the 29 previous inspectors inspected a second set of equivalent samples and 11 previously unavailable inspectors inspected their first set of samples.

Data were also collected on the age and length of service of the cooperating inspectors, and the amount of light under which each sample was inspected.

The results of this investigation are summarized below:

1. In relation to the grain standards, variability of initial estimates of all factors ranged from moderate to pronounced. Generally, the variability of estimates increased as the programmed values of WOOC, FM, and Damage

increased, whereas the variability of estimates of DHV decreased as the programmed value increased. The distribution of differences between initial and repeat estimates again disclosed considerable variability in relation to the criteria applied. All factors showed the same trends in variability as the programmed value of each increased as was found for initial estimates.

2. Initial estimates of FM and Damage tended to be below the programmed values, with this tendency increasing in magnitude as the value programmed for each increased. Estimates of DHV tended to be higher than programmed, with this tendency decreasing in magnitude as the value programmed increased. No consistent trend of over- or under-estimation of WOOC was found. This lack of consistency of estimation of WOOC was partially accounted for by the finding that the inspectors tended to pick Yellow Berry (YB) kernels as WOOC, and this tendency increased as the amount of YB in the sample increased. The magnitude of underestimation of Damage was found to be primarily related to the percentage of sick or black germ damage programmed. The differences between initial and repeat estimates of all factors except DHV showed no consistent tendency for repeat estimates to be either above or below initial estimates. Repeat estimates of DHV tended to be lower than initial estimates, with this tendency decreasing as the value of DHV programmed increased. The consistency of the relationship between initial and repeat estimates of the four factors by the same inspector ranged from strong for DHV and FM to moderately weak with pronounced individual differences in this ability for WOOC and Damage.

3. No consistent relationship was found between the inspectors estimates of any one factor and the same inspectors estimates of the other three factors investigated for any of three measures of error. Although some significant relationships were found, no overall picture of relationship between estimates of the different factors was achieved.

4. Neither age nor length of service of the inspector was found to be consistently related to the direction, magnitude, or variability of errors of estimation for the factors investigated, although some isolated relationships were found. However, the station at which the inspector was located was related to the amount of error of estimation for all factors except WOOC.

5. No relationship was found between the amount of light under which the samples were inspected, and the inspectors estimates of any of the factors investigated.

6. The reliability of both initial and repeat estimates of grade and subclass of the composite sample, when evaluated in relation to the grain standards, revealed that 40 per cent of the initial estimates placed the sample improperly, and that 42.5 per cent of the repeat estimates changed the grade and/or subclass of the sample, while 20.5 per cent of the samples were evaluated improperly on grade and/or subclass on both occasions.

7. Three possible explanations of the lack of reliability and accuracy of these determinations were offered. (a) The samples programmed could be considered as being more complex and for the most part of lower grade than the samples normally inspected. Thus the experimental samples could be considered as constituting an unfair test of the inspector's ability. (b) The variability of initial estimates, and the consistency of the relationship between initial and repeat estimates of DHV and FM, and the differences in amount of error associated with station, suggest thatthe inspectors, due to different training, personal interpretation of the grain standards, and the possible establishment of local norms for the interpretation of the grain standards, are all able to make these determinations, and the only problem is removing the differences between inspectors in the norms used. (c) All of the data except the consistency of the relationship between initial and repeat estimates of DHV and FM,

can be interpretated as indicating that the determinations required by the grain standards are so defined that the inspectors are incapable of consistently and accurately responding to the limited cues available to make these determinations. The definitions given in the grain standards are not highly specific in the characteristics whereby the various determinations are to be made, and in some instances impose arbitrary cutting points on essentially continuous gradations.

All three explanations, especially the last two, have merit, and on the basis of the present data, none can be disregarded. The relative merit of each will have to await completion of further research.

ACKNOWLEDGMENTS

The writer would like to take this opportunity to express his thanks and appreciation to the following individuals and organizations without whose cooperation this study could not have been successfully completed.

The writer is grateful to the Kansas State Agricultural Experiment Station for providing the funds for the research, and Dr. Max Milner for providing the original idea for the study. He is greatly indebted to Professors Ernest Mader and Howard Wilkins of the Kansas State University Agronomy Department and Professor Karl F. Finney of the Flour and Feed Milling Industries Department for their aid and advice as subject matter experts in the preparation of the experimental samples, and Curtis Adams, Kenneth Cross, and Robert McCay who assisted in the collection of the data.

Special appreciation is expressed to Arthur Cretin, Director of the Kansas Grain Inspection Department and Walter Sanderson, Director of the Missouri Grain Inspection Department and to the 40 licensed grain inspectors who so freely and willingly gave of their time and effort to cooperate in the study and inspect the experimental samples.

The writer would like to express his sincere appreciation of the hours of assistance, the expression of encouragement, and the patience extended him by Dr. Don Trumbo and the other members of the Department of Psychology.

In closing, the writer would like to acknowledge the assistance of his wife in preparing the manuscript, and also her help and encouragement throughout the study.

REFERENCES

Guilford, J. P. Psychometric methods. (2nd ed.) New York: McGraw-Hill, 1954.

Johnson, D. M. The psychology of thought and judgment. New York: Harper, 1955.

U. S. Department of Agriculture, Agricultural Marketing Service. Service and regulatory announcements, No. 148. (1941 rev.) Washington: U.S. Government Printing Office, 1941.

U. S. Department of Agriculture, Agricultural Marketing Service, Grain Division. Inspected receipts of wheat and rye. Chicago: 1959.

U. S. Department of Agriculture, Production and Marketing Administration, Grain Branch. Grain inspection manual. (1952 rev.) Washington: 1952.

APPENDICES

Appendix A

The formula for obtaining the 99 per cent confidence limits for the amount of DHV in each of the samples is as follows:

$$\text{Upper limit of DHV} = a + \left[\left[(\bar{x}_2 \cdot b) + [s_{.10}(\sigma_2 \cdot b)]\right] - \left[(\bar{x}_1 \cdot a) - [s_{.10}(\sigma_1 \cdot a)]\right]\right]$$

$$\text{Lower limit of DHV} = a - \left[\left[(\bar{x}_2 \cdot b) - [s_{.10}(\sigma_2 \cdot b)]\right] - \left[(\bar{x}_1 \cdot a) + [s_{.10}(\sigma_1 \cdot a)]\right]\right]$$

a = per cent of DHV programmed in sample

b = per cent of YB programmed in sample

$\bar{x}_1$ = mean amount of YB contamination in DHV population (1.15%)

$\bar{x}_2$ = mean amount of DHV contamination in YB population (0.915%)

σ_1 = standard deviation of YB contamination in DHV population (0.57%)

σ_2 = standard deviation of DHV contamination in YB population (0.57%)

$s_{.10}$ = standard-score value cutting off 10 per cent of the area under each tail of the normal curve.

The assumptions underlying the computation of the confidence limits are (1) that the distribution of the percentages of contamination found in each population, if an infinite number of samples were drawn would be normal, and (2) that the percentage of contamination found in a sample from one population is independent of the percentage of contamination found in any sample drawn from the other population.

The formulas used in the computation of the upper and lower limits of DHV are weighted for the amount of each of the populations in each sample because the larger the amount of each population, the greater the amount of contamination that can occur. The value of s used is determined by the fact that the probability of the joint occurrence of a high (low) amount of DHV contamination in the YB population and a low (or high) amount of YB in the DHV population is the product of the two independent probabilities. Since the levels of the joint

occurrence of a high and a low amount of contamination in the two independent populations was fixed at .01 in the beginning, the probability of occurrence of a high or a low amount of contamination in each of the two independent populations had to be the $\sqrt{.01}$ or .10. The length of the confidence interval remained constant over all samples, but changed in relation to the programmed value of DHV because the value of the two weighting factors in the formulas changed from sample to sample, but their sum remained constant. The values of the 16 confidence intervals are given in the table below:

Table 15

Values of 99 per cent Confidence Limits for 16 Programmed Percentages of DHV

Per Cent DHV Programmed	Upper Confidence Limit in Per Cent	Lower Confidence Limit in Per Cent
12	13.396	11.938
16	17.315	15.854
19	20.252	18.793
23	24.170	22.710
28	29.066	27.607
33	33.963	32.504
37	37.881	36.417
42	42.778	41.317
51	51.592	50.132
57	57.468	56.008
64	64.323	62.864
74	74.117	72.657
78	78.034	76.574
83	82.931	81.471
89	88.807	87.347
94	93.704	92.244

Appendix B

Table 16

Summary of Mean, Standard Duration and Range of Initial Estimates of Four Factors Investigated

DHV					Wooc					FM					Damage				
%	mean	S.D.	Range	N	%	mean	S.D.	Range	N	%	mean	S.D.	Range	N	%	mean	S.D.	Range	N
12	18.9	8.8	2 - 43	32	0	0.3	0.9	0 - 4	40	0	0	0	0 - 0.2	40	0	0.1	0.3	0 - 1.5	39
16	24.9	9.2	9 - 45	39	0	0.1	0.4	0 - 2	37	0	0	0	0 - 0.1	37	1	0.8	0.6	0 - 3.0	37
19	24.5	6.7	15 - 41	37	1	1.9	1.3	0 - 5	39	0.5	0.5	0.2	0.1 - 1.0	39	2	2.1	0.8	0.6 - 4.0	39
23	28.9	8.6	9.5 - 53	39	2	2.7	1.7	0 - 8	40	0.5	0.4	0.1	0 - 0.7	39	2	2.0	1.3	0.7 - 7.0	39
28	35.7	8.0	25 - 55	38	3	3.2	1.3	0 - 8	38	1	0.9	0.3	0.1 - 2.0	39	3	3.2	1.3	0 - 9.1	40
33	40.6	8.5	22 - 65	40	3	3.5	1.3	0 - 7	38	1	1.0	0.6	0 - 4.1	39	4	3.2	1.3	0.2 - 6.0	39
37	43.2	7.9	29 - 72	37	4	4.8	2.2	0 - 9.5	39	1.5	1.3	0.3	0 - 2.0	39	4	3.8	1.0	2.0 - 6.1	38
42	47.8	7.0	34 - 65	39	5	5.6	2.1	1 - 12	39	2	1.7	0.4	0 - 2.1	38	5	5.0	1.6	2.3 - 11.0	39
51	55.0	11.7	24 - 76	39	6	6.9	3.0	0 - 20	39	2	1.7	0.4	0.2 - 2.1	40	6	4.7	1.6	1.4 - 8.7	37
57	60.3	7.4	30 - 72	38	7	7.3	1.8	3 - 12	39	2.5	2.2	0.6	0.1 - 2.8	39	7	6.6	1.5	1.0 - 11.5	39
64	68.7	7.4	46 - 80	38	8	8.1	3.3	0 - 20	40	3	2.5	0.6	0 - 3.2	38	8	6.4	1.6	2.6 - 10.0	38
74	78.5	7.7	60 - 95	38	8	8.1	2.0	3 - 13	39	3	2.8	1.0	0.1 - 5.0	40	10	7.6	3.0	1.8 - 17.0	40
78	81.2	4.6	75 - 92	39	9	10.7	4.4	3 - 30	39	4	3.7	0.9	0.4 - 7.2	39	12	6.8	3.2	1.9 - 12.8	38
83	83.8	3.8	72 - 94	36	10	10.6	2.5	3 - 16	39	5	4.6	0.7	2.8 - 6.0	37	13	9.9	3.5	1.8 - 18.0	39
89	90.8	3.3	81 - 98	37	10	11.7	2.7	5 - 18	37	5	4.5	0.6	3.0 - 6.0	39	15	12.0	3.5	4 - 19.4	39
94	93.7	3.2	84 - 98	40	11	11.2	1.8	7 - 16	39	6	5.2	0.9	1.3 - 6.1	39	16	13.3	3.4	3.8 - 23.6	40

RELIABILITY OF SUBJECTIVE JUDGMENTS IN THE INSPECTION OF HARD RED WINTER WHEAT

by

CALVIN KELLY ADAMS

B. S., Kansas State University, 1958

AN ABSTRACT OF A THESIS

submitted in partial fulfillment of the

requirements for the degree

MASTER OF SCIENCE

Department of Psychology

KANSAS STATE UNIVERSITY
OF AGRICULTURE AND APPLIED SCIENCE

1960

This report presents a psychophysical investigation of the inter- and intra-inspector reliability and accuracy of four subjective determinations made in the inspection of Hard Red Winter Wheat by practicing licensed grain inspectors under field conditions. The criteria used to evaluate the accuracy and reliability of these determinations was the standards given in the United States Grain Standards Act. Equivalent sample sets of sixteen samples each, containing accurately programmed values of Dark, Hard, and Vitreous (DHV), Wheats of Other Classes (WOOC), Foreign Material (FM), and Damaged (Damage) kernels, were prepared. These sample sets were then administered to a total of 40 practicing licensed grain inspectors at nine different inspection points throughout Kansas and Missouri on two different occasions. On the first occasion 29 inspectors completed sample sets, on the second occasion, approximately two months later, 21 of the 29 previous inspectors inspected a second set of equivalent samples and 11 previously unavailable inspectors inspected their first set of samples.

Data were also collected on the age and length of service of the cooperating inspectors, and the amount of light under which each sample was inspected.

The results of this investigation are summarized below:

1. In relation to the grain standards, variability of initial estimates of all factors ranged from moderate to pronounced. Generally, the variability of estimates increased as the programmed values of WOOC, FM, and Damage increased, whereas the variability of estimates of DHV decreased as the programmed value increased. The distribution of differences between initial and repeat estimates again disclosed considerable variability in relation to the criteria applied. All factors showed the same trends in variability as the programmed value of each increased as was found for initial estimates.

2. Initial estimates of FM and Damage tended to be below the programmed

values, with this tendency increasing in magnitude as the value programmed for each increased. Estimates of DHV tended to be higher than programmed, with this tendency decreasing in magnitude as the value programmed increased. No consistent trend of over- or under-estimation of WOOC was found. This lack of consistency of estimation of WOOC was partially accounted for by the finding that the inspectors tended to pick Yellow Berry (YB) kernels as WOOC, and this tendency increased as the amount of YB in the sample increased. The magnitude of underestimation of Damage was found to be primarily related to the percentage of sick or black germ damage programmed. The differences between initial and repeat estimates of all factors except DHV showed no consistent tendency for repeat estimates to be either above or below initial estimates. Repeat estimates of DHV tended to be lower than initial estimates, with this tendency decreasing as the value of DHV programmed increased. The consistency of the relationship between initial and repeat estimates of the four factors by the same inspector ranged from strong for DHV and FM to moderately weak with pronounced individual differences in this ability for WOOC and Damage.

3. No consistent relationship was found between the inspectors estimates of any one factor and the same inspectors estimates of the other three factors investigated for any of three measures of error. Although some significant relationships were found, no overall picture of relationship between estimates of the different factors was achieved.

4. Neither age nor length of service of the inspector was found to be consistently related to the direction, magnitude, or variability of errors of estimation for the factors investigated, although some isolated relationships were found. However, the station at which the inspector was located was related to the amount of error of estimation for all factors except WOOC.

5. No relationship was found between the amount of light under which the

samples were inspected, and the inspectors estimates of any of the factors investigated.

6. The reliability of both initial and repeat estimates of grade and subclass of the composite sample, when evaluated in relation to the grain standards, revealed that 40 per cent of the initial estimates placed the sample improperly, and that 42.5 per cent of the repeat estimates changed the grade and/or subclass of the sample, while 20.5 per cent of the samples were evaluated improperly on grade and/or subclass on both occasions.

7. Three possible explanations of the lack of reliability and accuracy of these determinations were offered. (a) The samples programmed could be considered as being more complex and for the most part of lower grade than the samples normally inspected. Thus the experimental samples could be considered as constituting an unfair test of the inspector's ability. (b) The variability of initial estimates, and the consistency of the relationship between initial and repeat estimates of DHV and FM, and the differences in amount of error associated with station, suggest that the inspectors, due to different training, personal interpretation of the grain standards, and the possible establishment of local norms for the interpretation of the grain standards, are all able to make these determinations, and the only problem is removing the differences between inspectors in the norms used. (c) All of the data except the consistency of the relationship between initial and repeat estimates of DHV and FM, can be interpretated as indicating thatthe determinations required by the grain standards are so defined that the inspectors are incapable of consistently and accurately responding to the limited cues available to make these determinations. The definitions given in the grain standards are not highly specific in the characteristics whereby the various determinations are to be made, and in some instances impose arbitrary cutting points on essentially continuous

gradations.

All three explanations, especially the last two, have merit, and on the basis of the present data, none can be disregarded. The relative merit of each will have to await completion of further research.