**On the Triangle Test with Replications**

Joachim Kunert and Michael Meyners

Fachbereich Statistik, University of Dortmund, D-44221 Dortmund, Germany
E-mail: kunert@statistik.uni-dortmund.de
E-mail: meyners@statistik.uni-dortmund.de

**Abstract**

We consider the triangle test with replications, i.e. each assessor is asked repeatedly. A commonly used test statistic for this situation is the sum of all correct assessments, summed over all assessors. Several authors (e.g. o'Mahony, 1982, Brockhoff and Schlich, 1998) argue that the binomial distribution cannot be used to analyse this kind of data. Brockhoff and Schlich (1998) propose an alternative model for the triangular test with replicates, where the assessors have different probabilities to correctly identify the odd sample even if the products are identical.

Although we agree that assessors will have different probabilities of correct assessment if there are true differences, we do not think that Brockhoff and Schlich's model makes sense under the null hypothesis of equality of treatments. We show that all assessments are independent and have success probability 1/3, if the null hypothesis is true and the experiment is properly randomized and properly carried out. This implies that the sum of all correct assessments is binomial with parameter $p = 1/3$. Therefore the usual test based on this sum and the critical values of the binomial distribution is a level $\alpha$ test for the null hypothesis of equality of the products, even if there are replications.

## 1 Introduction

Because they are carried out easily and provide a simple and straightforward analysis, triangle tests are widely used in sensory analysis. In a triangle test, an assessor gets presented three samples which come from two products. Two of the samples are from the same product, the third sample is from the other. The assessor is asked to identify which is the odd sample. He / she is asked to make a choice, even if no difference is perceived.

With triangle tests, many observations may be needed to get sufficiently high power to show significance if there are only small differences between the products. There may be not enough assessors available to have the desired number of assessments. Then it is convenient to let each assessor test repeatedly. Such experiments are commonly analysed as if there were no replications, i.e. as if all assessments came from different assessors. Using the notation of Brockhoff and Schlich (1998) let $n$ denote the number of assessors each of which performed $k$ replications. If $m$ denotes the number of assessments, then $m = n\,k$. We say that the assessor had a success in a given replicate, if the right answer is given, i.e. the sample that differs from the two others is identified. The number $x_i$ of correct assessments of the $i$-th assessor is calculated, and these numbers are added over the assessors to get the number $x$ of all correct assessments. Then in this naïve approach $x$ is compared to the critical value of the binomial distribution with parameters $m$ and 1/3.

It has been argued that the binomial distribution is not adequate for the evaluation of such triangle tests with replications (see e.g. o'Mahony, 1982, or Brockhoff and Schlich, 1998). If an assessor is able to perceive the difference between the products and therefore gives a right answer once, then he will most likely perceive it again in a second replicate. If, however, another assessor is not sensitive enough to perceive the difference between the products in one trial, then he will most likely not perceive it in a second trial. Therefore the assessors have different probabilities of successes and $\Sigma_i x_i$ is not binomially distributed.

For discrimination tests with replications, Brockhoff and Schlich (1998) therefore propose to adjust the number $m$ of observations according to some variability criterion. This criterion depends on the overdispersion observed in the data. The larger the overdispersion is, the more the number of observations gets reduced.

We can not fully agree to these arguments. We show that the naïve binomial test can also be used in this situation. Under the null hypothesis of equality of products and under proper randomization of the design the number of correct assessments is binomial with success probability 1/3. This implies that (under the null hypothesis) the observations can be treated as if they were all produced by different assessors and there were no replications. Therefore, the naïve test which compares the number of correct assessments to the 1-$\alpha$ critical value of the binomial distribution with parameters $m$ and 1/3 is a level $\alpha$ test for this null hypothesis.

For the situation that there are differences between the products, we suggest an alternative model, which is a variant of Brockhoff and Schlich's (1998) model. If the products are equal, then in our model all assessors have the same success probability 1/3.

Our considerations indicate that Brockhoff and Schlich's (1998) method is too conservative. We recalculate two of the artificial examples in Brockhoff and Schlich (1998) which, we think, do give strong hints on product differences.

## 2 Model assumptions

As a first step, we assume that there is no sensory difference between the two products A and B. Then an assessor has a certain strategy to decide which of the three samples he selects as the odd one. This strategy may be random or systematic. For our considerations, there is one basic assumption: We assume that under the null hypothesis of product equality, the response of the assessor is independent of the order in which the products are presented. Call this assumption H (because it is valid only under the null hypothesis). This assumption is plausible, if the experiment is carried out properly. This includes, for instance, that the two products presented are of equal appearance, temperature, etc.

The experimental design is a random process which determines which of the six possible orderings AAB, ABA, BAA, ABB, BAB, BBA is presented to the assessor. Assume the process which determines the ordering is such that each of the six possible orderings is presented with equal probability. Due to assumption H, the probability of a correct selection of the odd product then is 1/3.

Now assume a second presentation is made to the same assessor. It is clear that the second choice of the same assessor is not independent of the first choice. It is possible, for instance, that the assessor always changes position, that is he chooses another position at the second presentation. It is also possible that another assessor may always select the same position.

However, we still have assumption H. Assume the ordering for the second presentation is randomized independently of the first presentation, such that it gives equal probability to all six possible orderings. Whatever strategy the assessor might apply, under assumption H the probability that the odd sample is placed on the position that the assessor chooses, remains 1/3 for the second trial, independent of the outcome of the first evaluation. Note that this holds if we do a third, forth, ... replicate, and it remains true both if we always have the same assessor and if the assessor is replaced by somebody else after some trials. It only is necessary that there are no sensory differences between the products. Therefore, under assumption H we have the following result: If there are $m$ presentations and the ordering of the products is randomized independently for each presentation, then the total number $x$ of correct guesses follows a binomial with parameters $m$ and $p = 1/3$. This remains true, whether or not we have replicates. It is the number of assessments that counts.

Usually, we rely more on a result that was produced by 100 assessors, each of which made one choice, than on a result that was produced by just one assessor who made 100 choices. However, a significant result that was derived from just one assessor also controls the type I error, that our test might indicate sensory differences which are not really there. The null hypothesis implies assumption H. Even if we have only one assessor and he / she gives a correct answer in significantly more than one third of the assessments, then sensory differences have been proven.

The number of assessors gets important for the power of the procedure. If there are 100 assessors, and only 35 of them correctly identified the odd sample, then we have good reasons to believe that there is a very small difference between the products, if any. If we have just one assessor, who correctly identified in only 35 out of 100 replicates, then we can only be sure that the difference between the products is too small for this assessor.

Whenever a sensory difference is present between the two products, then we can assume that there are "good" assessors, who do experience the difference, and "poor" assessors who do not. Let us consider the extreme case that exactly one half of our assessors will always give the right answer, while the other half will only guess. Assume that we do a test with significance level 5% and $m$ = 100 assessments, and assume there are two possible ways to do the experiment.

Case 1: We have just one assessor who gives 100 answers.

Then we have probability 1/2 that this one assessor is "good", in which case we will get 100 correct answers. There also is probability 1/2 that he / she is "poor", which will lead to a number of correct answers that is a binomial with parameters $m$ = 100 and $p$ = 1/3. Therefore, the probability of a significant result at the 5%-level is 1 if the assessor is "good" and 0.05 if the assessor is "poor", giving an overall probability of 0.525 of rejecting the null hypothesis.

Case 2: We have 100 assessors each giving exactly one answer.

Then for each assessor, we have probability 1/2 that he / she is good. With a good assessor the answer is correct with probability 1. We also have probability 1/2 that the assessor is poor, in which case the answer is correct with probability 1/3. In all, each assessor's answer has probability 2/3 to be correct. Therefore the number of correct answers is a binomial with parameters $m$ = 100 and $p$ = 2/3. This implies that there is a probability of more than 99 % to observe more than 42 correct answers. Since 42 is the critical value of the triangle test with 100 assessments, we therefore have a probability of more than 99% of correctly rejecting the null hypothesis.

After this artificial example, we also extend our model to the alternative that there are product differences. We assume that there are two different groups of assessors, those who are able to perceive the difference between the two given products and those who do not perceive the difference between these two products because it is too small for them. Since the following considerations also hold for other discrimination tests, let some more general $c$ be the probability to succeed by chance (e.g. $c$ = 1/3 for the triangle test or $c$ = 1/2 for the duo-trio test). Furthermore let $p_i$ be the probability for assessor $i$ to have a success, $i = 1,\ldots,n$.

The proportion of assessors in the first group is denoted by $\gamma$, where $0 \leq \gamma \leq 1$. If there is no difference between the samples, then the first group is empty, i.e. $\gamma = 0$. In fact, we might define that no sensory difference exists if and only if $\gamma = 0$.

For each sensitive assessor, that is for each assessor in the first group, the probability of a success increases to a number $p_i = \pi_i + (1-\pi_i)c > c$. Here $\pi_i$ is the probability that assessor $i$ actually identifies the odd product (and not only guesses). We might consider $\pi_i$ to be a random variable (if we assume that the assessors are drawn from some superpopulation), but it is clear that $\pi_i > 0$ because it is a probability. In the examples in section 4, we assume that all $\pi_i = 1$.

We do not, however, generally assume that $\pi_i = 1$, because even those assessors who are able to experience the difference, might miss it with some presentations, e.g. due to random variation between the samples. If the difference between the products increases, then $\gamma$ as well as the $\pi_i$ will tend to 1.

Now assume that the assessors are drawn at random from some superpopulation, such that each assessor has a probability of $\gamma$ to come from the first group, and a probability of $1 - \gamma$ to come from the second group. Therefore the probability $p_i$ of assessor $i$ to succeed can be written as

$$p_i = \begin{cases} c & \text{with probability } 1-\gamma \\ \pi_i + (1-\pi_i)c & \text{with probability } \gamma \end{cases}$$

where $\pi_i$ is a realization of a positive random variable less or equal 1. We do not specify the distribution of the $\pi_i$. This distribution depends on the population of the assessors, on the difference between the products, etc. The distribution of the $\pi_i$ has to be modelled, it is not determined by the experimental design and the randomization. This is different from the distribution of $x$ under the null hypothesis.

For assessor $i$ we assume $p_i$ to be constant during all of his / her replications. This is reasonable only under the restriction that the number of replications is small enough that fatigue effects can be neglected.

## 3 Comparison to Brockhoff and Schlich (1998)

Brockhoff and Schlich (1998) propose a model with random assessor effects, i.e. we have

$$p_i = \pi + \varepsilon_i, \quad i = 1,\ldots,n,$$

where $\pi$ is the average probability of an assessor $i$ to succeed, and the average is taken over a population of possible assessors. If there is no difference between the products, then $\pi$ is 1/3 for the triangle test. The random variable $\varepsilon_i$ has zero mean and an unknown variance. It does not vanish if there is no difference between the products.

Note that with these assumptions, if there is no difference between the products, then since $\pi = 1/3$ there will be some assessors $i$ with $p_i < 1/3$. That is, the model of Brockhoff and Schlich (1998) implies that for some assessors the probability of a correct result gets less than what we would expect from pure guessing. We do not think that this is reasonable if the data come from a properly designed experiment: If there is no sensory difference between the products, how should an assessor manage to systematically get the wrong sample? There are two possible ways, both of which can be excluded by the design of the experiment.

First option: The assessor might find out which sample comes from which product by some other means than the sensory difference, for instance the products are identified in such a way that the assessor can solve the code. It is clear that a properly designed experiment will exclude such possibilities. We assume here that the experiment is run in such a way that assumption H holds.

Second option: The assessor has a strategy which has a tendency to select a position where the experimenter did not put the odd sample. Experience shows that most assessors have a preference for a certain position, when they perceive no difference between the products. If the experimenter has a tendency to place the odd sample preferably on one of the positions that this assessor does not prefer, then the assessor has a probability of less than 1/3 of guessing correctly. This, however, can be avoided by randomization. If the odd sample is randomized to go to each position with equal probability then, under H, each assessor will have probability 1/3 of guessing right.

Presumably, the model of Brockhoff and Schlich (1998) was intended to model correlations between the responses of one assessor. It follows from their model that if one assessor gave a correct answer in the first replicate, then he has a higher probability of giving a correct answer in the second replicate. This is because assessors who gave a correct answer have a higher probability to have a large $p_i$.

However, under assumption H, correlations between the answers can be avoided by randomization. In fact, correlations between the answers are indications that either there are differences between the products or that a poor randomization has been used. An example of a poor randomization is if the experimenter randomizes just once for each assessor and uses the same presentation in every run. Then an assessor who has a tendency towards a given position, and who guessed right in the first replicate, has a higher probability to guess right in the second replicate, too. He only has to stick to his position. With independent randomization, however, the fact that an assessor guessed right in the first replicate has no influence on his chance to guess right in the second replicate, whatever strategy the assessor might have.

There is another randomization which is used frequently in triangle tests, but which might cause correlations. Quite often, the randomization is not done independently, but in a way to make sure that if an assessor gets an ordering in the first replicate, then he gets another ordering in the second replicate. Such a randomization is recommended in e.g. the ISO-standard on the triangle test (ISO 4120, 1983). Now assume that an assessor uses the strategy to "change positions", i.e. if he opted for one position in the first trial, then he will use another position in the second trial. Then under the randomization proposed in the ISO-standard, the result of the second replicate is no longer independent from the outcome of the first. To see this, assume the assessor guessed right in the first replicate. He will choose another position in the second trial. The experimenter will not use the ordering he had in the first replicate, he will choose one of the five other possible orderings. Four orderings among these five have a different position for the odd sample, and two have the position chosen by the assessor. Therefore, the probability to guess right after a first correct guess becomes $2/5 > 1/3$. An assessor who sticks to his position, however, will have a smaller probability for a second success after a success in the first replicate. Such negative correlations between the successes cannot be modelled by Brockhoff and Schlich's (1998) model. Remember, however, that these correlations only occur if the randomization recommended in the ISO standard is applied.

Hunter (1996) suggests the number of replicates (if any) always to be a multiple of six. He proposes to randomize in such a way, that each assessor gets presented each ordering equally often. This is a special instance of the method criticised above. We do not consider this an appropriate randomization either. Only with independent randomization, the independence of the successes is guaranteed, provided the null hypothesis is true.

We have here a basic property of designed experiments that we think many experimenters do not sufficiently appreciate. At least under the null hypothesis of product equality, we can use randomization to introduce a simple distributional structure into the data. The idea of randomization was introduced by Fisher (e.g. 1935). The tool of randomization has been well examined under mathematical as well as practical aspects. The philosophy of randomization is explained in e.g. Bailey (1981).

Things get different if the null hypothesis is not true. The way in which the differences between the products influence the response has to be modelled. It cannot be explained by randomization theory. We think that in the case of the triangle test with replicates, a model as the one described in Section 2 is reasonable. Therefore, for power calculations we must take into account that there are replicates and we must model the distribution of the $\pi_i$. It is beyond the scope of this paper to deal with the problem of how power calculation should generally be done.

The aim of the paper is to point out that the naïve test, which pools the number of correct guesses, provides a valid test to show that there are significant differences between products. This is true if the data were derived from an experiment that was properly randomized, and that was also properly carried out such that assumption H can be justified. Since we do not know the power of the naïve test, it cannot, however, be used to show that there are no differences between products.

Let us return to the extreme situation that we have only one single assessor. Assume this person gives significantly more correct answers than what is possible by chance. Corresponding to what we have said so far, it is reasonable to decide that the two products under consideration differ from each other. However, if we have a group of assessors, then we cannot simply take the assessor with the highest number of correct guesses and do a triangle test based on his / her results only. If we have e.g. 100 assessors all of which do 10 evaluations, then we can expect that there are about 2 among them who have 7 or more correct guesses, even if there is no difference between the products. The analysis must pool all the assessors in the trial.

## 4 Examples

We revisit Examples 2 and 3 of Brockhoff and Schlich (1998). Both examples concern the triangle test, so again $c = 1/3$. Example 2 gives artificial data of a triangle test with $n = 12$, $k = 4$ and $x = 24$. The naïve test therefore gives a significant difference between the products at the five percent level. We look at the data a bit closer, to see why this is reasonable with data like this. There are 3, 2, 2, 2 and 3 assessors with 0, 1, 2, 3 and 4 correct answers, respectively. Now assume that there is no difference between the products. We might want to carry out the $\chi^2$ - goodness-of-fit-test to see whether the numbers $x_i$ of correct guesses are binomial with parameters 4 and 1/3. The computations are given in Table 1.

| | Number $j$ of correct results | | | | | |
|---|---|---|---|---|---|---|
| | 0 | 1 | 2 | 3 | 4 | sum |
| $P_j = Prob(x_i=j)$ | 16/81 | 32/81 | 24/81 | 8/81 | 1/81 | 1 |
| $n\ P_j$ | 2.37 | 4.74 | 3.56 | 1.18 | 0.15 | 12 |
| observed $x_i=j$ | 3 | 2 | 2 | 2 | 3 | 12 |
| $\frac{(nP_j - observed)^2}{nP_j}$ | 0.17 | 1.58 | 0.68 | 0.55 | 54.15 | **57.13** |

**Table 1:** Calculation of the $\chi^2$-statistic for the data from Brockhoff and Schlich, Example 2.

Note that the expected numbers in the cells are too small to assume that the $\chi^2$-statistics is distributed according to a $\chi^2$-distribution with 4 degrees of freedom. However, a calculated $\chi^2$-statistic of 57.13 is very large. It is obvious that the large size of the statistic is due to the 3 persons that succeeded in all 4 replications. If there was no difference between the products, then we would expect less than 1 assessor with four correct guesses in six experiments of this size. As argued earlier we think that with proper randomization, non-validity of the binomial distribution can only be explained if the products differ from each other. For simplicity, we assume all $\pi_i = 1$, that is every sensitive assessor, who can experience the difference at least once, gets it right in every replicate. Then the estimated proportion of sensitive assessors is (3-0.15)/12 = 23%. The method by Brockhoff and Schlich (1998) does not lead to the identification of a difference.

We now turn to example 3 of Brockhoff and Schlich, with $n$ = 100 assessors, $k$ = 3 replicates and $x$ = 112 correct answers. As these authors point out, “Anna Sens” wants to show that there is no difference between the products. The naïve test gives a difference at the 10% level.

As before we carry out a $\chi^2$ - goodness-of-fit-test to examine the data. The results are given in Table 2.

| | Number $j$ of correct results | | | | |
|---|---|---|---|---|---|
| | 0 | 1 | 2 | 3 | sum |
| $P_j = Prob(x_i=j)$ | 8/27 | 12/27 | 6/27 | 1/27 | 1 |
| $n P_j$ | 29.6 | 44.4 | 22.2 | 3.7 | 99.9 |
| observed $x_i=j$ | 34 | 33 | 20 | 13 | 100 |
| $\frac{(nP_j - observed)^2}{nP_j}$ | 0.65 | 2.93 | 0.22 | 23.38 | **27.18** |

**Table 2:** $\chi^2$ - goodness-of-fit-test for the data from Brockhoff and Schlich, Example 3.

Here the numbers are sufficiently large to do a $\chi^2$-test of significance. Then the result is highly significant, the corresponding $\chi^2$-distribution with 3 degrees of freedom gives a p-value of less than $10^{-5}$. As in the previous example this comes from the persons that guessed correctly in all replications. Once again, assume $\pi_i = 1$. Then if we subtract the 4 persons with three right guesses that we should expect under equality of the products, we estimate that there are 9 consumers who have really perceived the difference in all 4 trials. So we estimate that there is a difference of the products which is perceived by 9% of the consumers. Note that this is totally different from the conclusion of Brockhoff and Schlich (1998) who claimed that there is no difference between the products. In fact, we say the data give strong hints that there is a perceivable difference between the products. It may be argued that 9% of the consumers is too small a proportion for Anna Sens to worry about. However, if the experiment was run with 300 assessors, each of which is testing only once, then 9 % sensitive assessors would lead to a probability of 80% to identify the difference between the products. So, obviously, 9 % sensitive assessors is a margin for which Anna Sens has to expect a significant result with an experiment of this size.

Finally, we give an additional artificial example to illustrate why we do not regard the method of Brockhoff and Schlich (1998) as appropriate for difference tests in properly randomized experiments. We consider a somewhat extreme situation. Let us assume two consumers (i.e. $n$ = 2) that carry out $k$ = 100 replications of a triangle test under ideal conditions, neglecting any fatigue effects etc. Suppose one of the consumers succeeds in 34 of his trials. This is rather close to what we would expect if he can experience no differences between the products. However, the other assessor succeeds in all 100 replications. We test for product differences using the method of Brockhoff and Schlich (1998). The results of all single steps are listed in Table 3.

| $\hat{p}_1$ | $\hat{p}_2$ | $\hat{p}$ | $V_p$ | $\delta$ | $\delta_{\max}$ | $\delta_{\text{cor}}$ | $\hat{\sigma}^2$ |
|---|---|---|---|---|---|---|---|
| 0.34 | 1 | 0.67 | 0.2178 | 98.51 | 100.75 | 195.52 | 100 |

**Table 3:** Intermediate results for the method of Brockhoff and Schlich in the new example.

Since $\delta_{\text{cor}}$ is larger than $k$ we determine $\hat{\sigma}^2$ = $k$ = 100. Thus we adjust our number $nk$ of observations according to $\frac{nk}{\hat{\sigma}^2} = \frac{nk}{k} = n = 2$ and the number of successes according to $\frac{x}{\hat{\sigma}^2} = \frac{x}{k} = 1.34$. Following the suggestions we round this value to 1 and thus the p-value for the difference-test is just the probability to observe one or two successes from a *bin(2,1/3)*-distribution, which is 5/9 = 0.56, so we have no significance at all. However, it is quite clear that these products are different from each other.

## 5 Conclusions

The ideas of the paper are outlined by restricting on the triangle test. This is made to simplify the notation and to clear up our arguments. However, our considerations also hold for other discrimination tests like the e.g. duo-trio test.

We show that with a properly randomized experiment, then under the null hypothesis the number of successes in a triangle test is a binomial, even if there are replications from the single assessors. The basic consequence is that the naïve test, which pools all the successes of all assessors is a level $\alpha$ test.

We do not think that Brockhoff and Schlich's (1998) method of handling overdispersion is appropriate for the triangle test. This is shown with several examples. The basic difference is that we claim that with a properly randomized design assessor heterogeneity can only happen if there are differences between the products. Therefore, assessor heterogeneity is in fact an indication that the products are different.

This demonstrates why designed experiments are analysed much easier than observational studies. In observational studies, overdispersion has to be modelled, even under equality of the products.

**Acknowledgement**

The authors wish to thank Per Brockhoff and Pascal Schlich for letting us see their paper in advance.

The research for this paper was supported by the Deutsche Forschungsgemeinschaft, Sonderforschungsbereich 475.

## References

BAILEY, R. A. (1981): A unified approach to design of experiments. *Journal of the Royal Statistical Society* **A 144**, 214 – 223.

BROCKHOFF, P. B. and SCHLICH, P. (1998): Handling replications in discrimination tests. *Food Quality and Preference* **9**, 303 – 312.

FISHER, R. A. (1935): *Design of Experiments*. Oliver and Boyd, Edinburgh (8th edition, 1966).

HUNTER, E. A. (1996): Experimental design. *In:* Næs, T. and Risvik, E. (ed.): *Multivariate analysis of data in sensory science*, Elsevier, Amsterdam, 37 - 69.

ISO 4120 (1983): Sensory Analysis - Methodology - Triangular Test.

O'MAHONY, M. (1982): Some assumptions and difficulties with common statistics for sensory analysis. *Food Technology* **36**, 75 – 82.