The design of replicated difference tests

Michael Meyners
Fachbereich Statistik, Universität Dortmund, D-44221 Dortmund, Germany
Tel.: +49 231 755 3181, Fax: +49 231 755 3454
E-Mail: michael.meyners@udo.edu

AND

Per Bruun Brockhoff
Department of Mathematics and Physics, The Royal Veterinary and Agricultural University, Thorvaldsensvej 40, DK-1871 Frederiksberg C, Denmark
Tel.: +45 35 28 23 61, Fax: +45 35 28 23 50
E-Mail: pmb@kvl.dk

Abstract

We show that adding replications in replicated difference test results in larger power and smaller variance when the number of assessors is fixed. On the other hand, when the number of total assessments is fixed, the power usually decreases and the variability increases whenever replications are considered instead of different assessors. The appropriate numbers of assessor needed to gain the same power respectively variability when replications are used will be given. It is shown that the number of assessors might indeed be reduced, but this has to be paid for by an increasing total number of assessments. We show that two key models, namely the mixture binomial and a corrected version of the Beta-binomial model, are quite similar with respect to the properties of interest. We provide tables from which, according to her/his requirements, the investigator might find an appropriate setting with respect to the number of assessors and replications.

KEYWORDS: difference tests, replications, experimental design, power, variability, mixture binomial model, Beta-binomial model

Introduction

To consider whether or not differences between two (food) products of similar kind occur in the consumer population, difference tests are frequently used. As far as each consumer is only asked once, there are no doubts with respect to an appropriate analysis of the respective experiment. However, since the use of many assessors might become quite expensive, in practice a tendency towards difference tests with replications can be found. In this case, each assessor is asked to perform the test k times, say, where $k > 1$. From an intuitive point of view, the analysis of such data should depend on the assessor heterogeneity, such that the final statistical decision accounts for the fact that we considered less assessors but let them replicate, cf. Brockhoff and Schlich (1998). However, for the decision about whether or not perceivable differences exist at all, Kunert and Meyners (1999) show that the simple binomial test using nk observations is still applicable as far as the experiment has been properly designed. Here, n denotes the total number of assessors in the panel.

In general we are not only interested in deciding whether or not there are differences at all. In application, we might not worry about a single consumer or maybe one in a hundred who might find the difference at least once in a while. On the other hand, it will definitely be of importance to the investigator whenever nearly all consumers will taste the difference in every second trial. However, to judge upon the number of perceivers will strongly depend on the number of replications. Meyners (2002) shows that for given values of n and k, very different assessor performances may lead to an identical total number of successes. As well, he proposes an estimate of the lower limit of the number of consumers within the trial that must have perceived the difference at least once in a while.

Furthermore, we are often interested in testing for similarity, or at least in claiming it whenever no differences have been apparent. In order to prove similarity, though, investigations have to be given to the power of the test. Therefore we will also stress this question here. Kunert and Meyners (1999) use a simple and extreme example to show that the power might strongly decrease as soon as the number of replications increases while the total number of assessments remains fixed. They conclude that replications may be used, but will lead to less power for the test for product differences as might be gotten from the same number of trials with different assessors. In general, this means that for a given number of assessments, the design with the least number of replications will be the most powerful. In this paper, we will show that this conclusion does only hold for a reasonable design of the experiment in terms of the number of assessments in relation to the effect size of interest. On the other hand, if the corresponding condition is fulfilled, we will prove this result more formally.

Brockhoff (2002) calculates the power using different underlying models by means of Monte-Carlo-integration methods. Independent from the model under consideration, it could usually be seen that the power decreases as the number of replications goes up for a fixed nk. Still, the results were surprising due to the small decrease of power that was observed in this study and due to the counter intuitive slight increase of power in some cases.

We will give some rather theoretical considerations that support these findings. For this purpose, we will re-state two models proposed by Brockhoff (2002) and derive some properties under these assumptions. From this, we will determine different combinations of n and k which result in identical power of the corresponding test. Finally, we will also stress the variability of the statistics of interest, for which we will derive combinations of n and k as well, such that the variability remains the same while using any of these combinations. From the corresponding tables we might decide whether an increase of assessors or an increase of the number of replications is more reasonable with respect to the variability of the estimators. With it, we might also account for different costs in acquiring additional assessors in comparison to saving some assessors but increasing the number of assessments.

Note that similar considerations have been given by Ennis and Bi (1998) and Bi and Ennis (1999a). However, like Brockhoff and Schlich (1998), they argue that the statistical test for the null hypothesis has to be adapted in case of replications. Hence it is clear that their considerations and tables respect for this assumption. We do not agree with this at all and propose the use of the simple binomial test even for replicated difference tests, as it was pointed out in Kunert and Meyners (1999). Therefore the results of the different studies are not directly comparable as is the case for the power results of Brockhoff (2002) and Bi and Ennis (1999b).

Model assumptions

For the theoretical considerations, we assume a quite general model. Let P_i be a random variable which gives the success probability for assessor i, say. Furthermore, let X_i be the number of successes of assessor i, say. Then we assume that X_i under the condition that $P_i = p_i$ is binomial distributed with parameters k and p_i, shortly $\mathbf{L}(X_i|P_i = p_i) = \mathbb{B}(k, p_i)$.

Considering the distribution of the P_i for $i = 1, ..., n$, we first of all assume that the experiment has been properly designed such that the outcomes of the different assessors are independent from each other (Kunert and Meyners, 1999). Furthermore, we assume an appropriate sampling from the overall population of interest. In that case, all P_i's are identically distributed with an arbitrary distribution. This distribution represents the spreading of different success probabilities within the population under consideration. Note that these conditions are not very strict since they can be fulfilled by an appropriate design of the experiment!

Finally, we assume that once a value has been chosen for any particular i, this value is fixed for the respective assessor. This assumption implies that we neglect possible fatigue or training effects of the assessors, i. e. we assume that an assessor will have the same success probability throughout all her / his replications. This might be questionable for quite large numbers of replications k, while we guess that this assumption can be justified for reasonable values of k.

We do not generally assume any particular distribution of the random variables P_i, $i = 1, ..., n$, but some restrictions can be easily derived from the problem under consideration. We find that the distribution of any P_i has to satisfy two conditions: Since in a properly designed experiment, the success probability of any assessor may not become smaller than the success probability derived by pure guessing π_0, say, the respective values should not fall below π_0 (Kunert and Meyners, 1999). Neither should it exceed 1, since any realization of P_i represents a probability. Since $\pi_0 = 0$ is not useful for any practical applications, without loss of generality we confine ourselves to distributions on the interval $[\pi_0, 1]$ where $0 < \pi_0 \leq 1$.

Mean and variance of the total number of successes

We now consider the mean and the variance of the overall number of successes within the experiment. To start with, we calculate the respective values for a single assessor i. Once the value of P_i is fixed to be p_i for this assessor, the mean and variance of the random variable X_i given that $P_i = p_i$ are well known to be

$$\begin{aligned} \mathbf{E}(X_i|P_i = p_i) &= kp_i \\ \mathbf{Var}(X_i|P_i = p_i) &= kp_i(1 - p_i). \end{aligned}$$

By basic probability calculus rules we find the unconditional expectation and variance to be

$$\begin{aligned} \mathbf{E}(X_i) &= \mathbf{E}\left[\mathbf{E}(X_i|P_i)\right] \\ &= \mathbf{E}(kP_i) \\ &= k\mathbf{E}(P_i) \end{aligned}$$

and

$$\begin{aligned}
\mathbf{Var}(X_i) &= \mathbf{Var}\left[\mathbf{E}(X_i|P_i)\right] + \mathbf{E}\left[\mathbf{Var}(X_i|P_i)\right] \\
&= \mathbf{Var}(kP_i) + \mathbf{E}\left(kP_i(1-P_i)\right) \\
&= k^2\mathbf{Var}(P_i) - k\mathbf{E}(P_i^2) + k\mathbf{E}(P_i) \\
&= k^2\mathbf{Var}(P_i) - k\mathbf{E}(P_i^2) + k\left(\mathbf{E}(P_i)\right)^2 - k\left(\mathbf{E}(P_i)\right)^2 + k\mathbf{E}(P_i) \\
&= (k^2-k)\mathbf{Var}(P_i) + k\mathbf{E}(P_i)(1-\mathbf{E}(P_i)).
\end{aligned}$$

Noting that the X_i are independently identically distributed, we can derive the respective values for the total number of successes X, say, within the experiment as

$$\begin{aligned}
\mathbf{E}(X) &= \mathbf{E}\left(\sum_{i=1}^{n} X_i\right) \\
&= n\mathbf{E}(X_1) \\
&= nk\mathbf{E}(P_1) \qquad (1)
\end{aligned}$$

and

$$\begin{aligned}
\mathbf{Var}(X) &= n\mathbf{Var}(X_1) \\
&= n(k^2-k)\mathbf{Var}(P_1) + nk\mathbf{E}(P_1)(1-\mathbf{E}(P_1)). \qquad (2)
\end{aligned}$$

Hence we find that the expectation and variance of X depend on the expectation and variance of the random variable P_1, which cannot be influenced by means of the design of the experiment. However, they depend as well on the numbers n and k, which indeed can be influenced by the experimental design. To be precise, the expectation of the total number of success depends, besides the distribution of P_1, only on the total number of assessments nk, i. e. the expected test statistic does not depend on whether or not the assessments have partly been derived from replications.

The variance of X also depends on the total number of assessments nk, as can be seen directly from the term at the right hand side of (2). However, the term on the left hand side depends on $n(k^2 - k)$ and therefore disappears for $k = 1$. Re-writing this as $nk(k-1)$ directly shows that for a fixed number of assessments nk this term monotonically increases with the number of replications k. Hence, independent of the distribution of P_1, the variance of the total number of successes increases with the number of replications as far as the total number of assessments remains constant.

Linking variability to the power of tests

Unfortunately, no direct link between the variability of the test statistic and the power of the test holds in general. Nevertheless, in this section we will give a link that will hold whenever the design of the experiment fits to the effect size in our terms, meaning that the power of the test is not smaller than 50%. If so, the median observed test statistic is not smaller than the critical value c, say. In case of a symmetrical distribution of the test statistic, the median is equal to the mean of the distribution, such that a power of more than 50% is equivalent to the mean of the test statistic being larger than c. In this case we might say that we would at least expect the rejection of the null hypothesis of the respective test.

To illustrate this, consider a triangle test (i. e. $\pi_0 = \frac{1}{3}$) and an arbitrary distribution of P_1 such that $\mathbf{E}(P_1) = 0.45$, while we fix $nk = 50$. Then we would expect 22.5 correct answers. Since the critical value is equal to 21 at a 5%-level, we therefore would expect to reject the null hypothesis of product equality. However, the variance of X is smaller whenever we have $n = 50$ and $k = 1$ in comparison to, e. g., $n = 25$ and $k = 2$ or $n = 10$ and $k = 5$. Running a large number of identical experiments, with $n = 50$ and $k = 1$ we might hope to observe a lot of trials with $x = 22$ or $x = 23$, which would lead to a rejection of the null hypothesis. On the other hand, in one of the latter cases, we will more likely observe values $x \leq 21$ as well, such that we cannot reject the null hypothesis. Of course, at the same time we will observe more larger values as, e. g., 25 or 26, however these have no additional use compared to a value of 22: in both cases the test decision is the same! Recalling the results from the section before concerning the variability, it can be seen that we indeed might loose power while using replications. This case is illustrated in figure 1. Note that the distribution is symmetric such that the mean is equal to the median.

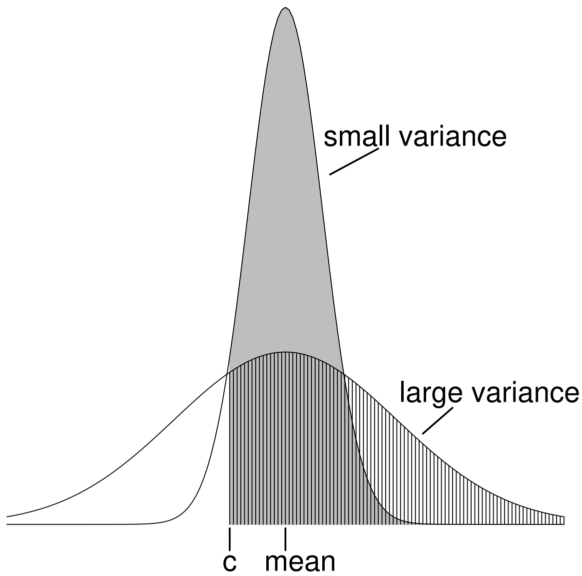

Figure 1: Power of the test for a mean slightly larger than the critical value c in case of small and large variance, respectively. The colored areas represent the power of the test.

This consideration holds to a larger extent the closer the expected value of X is to the critical value, whereas for an expected value of 30, say, in the upper example, the difference in power might be negligible. This is illustrated in figure 2, in which the loss of power due to the increased variability is given by the black area.

However, as stated before, this link does not hold in general. If we had considered $\mathbf{E}(P_1) = 0.40$ in our example, we would have only expected 20 correct answers and therefore we would expect not to reject the null hypothesis. Following the same argumentation as before, we find that in this case the power of the test increases with an increasing number of replications! The graphical representation of this case can be found in figure 3.

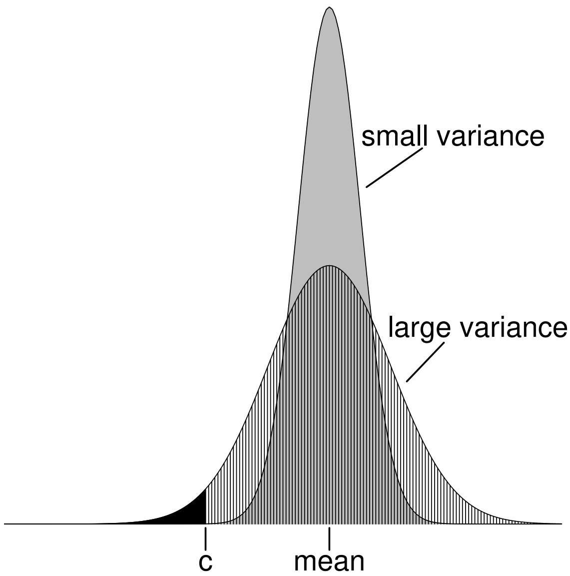

Figure 2: Power of the test for a mean much larger than the critical value c in case of small and large variance, respectively. The small black area at the left represents the loss of power due to the increased variability.

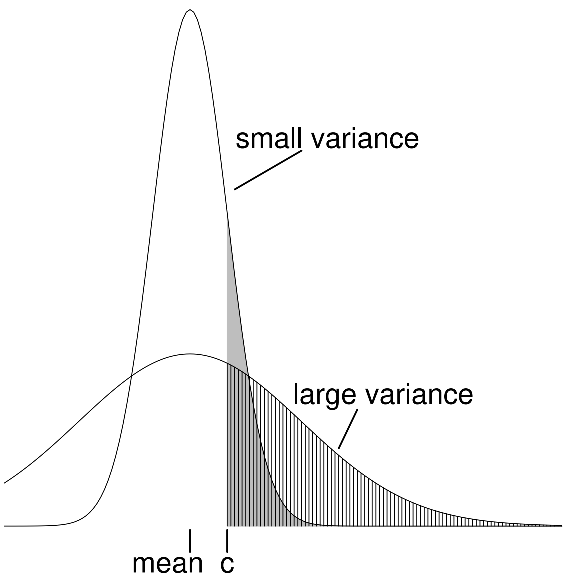

Figure 3: Power of the test for a mean smaller than the critical value c in case of small and large variance, respectively. The colored areas represent the power of the test.

More formally, stressing the asymptotic of the central limit theorem this can be seen as follows. The random variables $X_1, ..., X_n$ are assumed to be independently identically distributed. Denote the mean of X by μ, say, and the variance by σ^2, say. From the central limit theorem we know that $\mathbf{L}\left(\frac{X-\mu}{\sqrt{\sigma^2}}\right) \to \mathbb{N}(0,1)$ for $n \to \infty$, i. e. the distribution of X converges to the standard normal distribution. For the probability $\mathrm{P}(X \geq c)$ to reject the null hypothesis we find

$$\begin{aligned} \mathrm{P}(X \geq c) &= \mathrm{P}\left(\frac{X-\mu}{\sqrt{\sigma^2}} \geq \frac{c-\mu}{\sqrt{\sigma^2}}\right) \\ &\approx 1 - \Phi\left(\frac{c-\mu}{\sqrt{\sigma^2}}\right), \qquad (3) \end{aligned}$$

while Φ is the distribution function of the $\mathbb{N}(0,1)$. If the mean μ is larger than the critical value, the argument of the distribution function Φ is negativ and therefore decreases with a decreasing variability. Thus the probability to reject the null hypothesis increases, as we usually would hope. On the other hand, if $c - \mu > 0$, the argument of Φ is positive and therefore increases with an decreasing variance. In that case, the power of the test actually decreases with the variability. Note that due to the discreteness of X, in both cases the power may occasionally remain unchanged, since a small change of the argument of the corresponding discrete distribution function may result in an identical value. However, at least it will never behave in the opposite way to the one stated.

For a fixed number of assessors n, the variability of the total proportion of correct responses will decrease with an increasing number of replications k. This intuitive property will be formally proven in the appendix, while it may also be found under slightly different assumptions from, e.g., McCullagh and Nelder (1989). Knowing this, from the considerations given before it can directly be seen that the power of the corresponding test will increase whenever we have enough assessments in comparison to the effect size such that the power of the test is larger than about 50%.

Note that the power of the test and the variability of estimators aim at two different goals. While the variability is of interest mainly in order to get an as precise as possible estimator for, e. g., the mean effect size, the power of the test is particularly essential whenever we do not only aim at showing differences, but also at claiming similarity when the null hypothesis cannot be rejected. We want to stress that this is indeed a quite different task in comparison to only aiming at proving differences.

To illustrate the difference, we have to recall that statistical testing is an asymmetrical procedure. We more or less easily control the level α, which gives the probability to wrongly reject the null hypothesis and accept the alternative. On the other hand, the probability β to mistakenly accept the null hypothesis is much harder to determine and can usually only be influenced by changing the number of observations. In addition, the opposite of the null hypothesis has to be much more precisely defined than we usually would do. In our context this means that we will never be able to state that there are absolutely no differences between two products by means of a triangle test. This is due to the fact that for any fixed number of observations and level α, the error rate β will tend to 1 as soon as the effect size tends to π_0, i. e. the success probability by pure guessing. The only statement we might ever achieve is that the effect between the products under consideration is not larger than a given value γ, say. Given values of π_0, α and γ, we may then control the power of the test $1-\beta$ by using different numbers of observations.

As we have found from (2), whenever the total number of assessments nk is fixed, an increase of k debiting n will increase the variability. From what has been derived earlier in this section from (3), we find that the probability $1-\beta$ will decrease at the same time and therefore the power of the test goes down, given that the power has been larger than 50% before. This is also intuitively clear, since we have to rely on the few assessors only. In an extreme case, we might only prove that a particular assessor does not perceive a difference, but this will not be representative at all, cf. the extreme example in Kunert and Meyners (1999). However, the question arises whether we may all the same use less assessors but increase the total number of assessments instead.

This question has been addressed by Brockhoff (2002) in more detail. He showed that even for small numbers of assessor, the power of a test can become reasonably high whenever some replications are considered. We will stress his results from another point of view. To prove similarity, we should consider a small error rate β respectively a large power $1-\beta$. We propose that the power should not fall below a minimum of 90%, while 95% or 99% are to be preferred. (These values correspond to a significance level of 10%, 5% and 1%, respectively, in proving differences.) We consider a worst case scenario, assuming a maximal heterogeneity for a given effect size. This is given whenever the assessors are either non-perceivers with success probability π_0 or perceivers with success probability 1 in all replications. Under these assumptions, Brockhoff's (2002) tables 3 and 4 give the power of the corresponding binomial test for all combinations of $n \in \{5, 6, ..., 50\}$ and $k \in \{1, 2, 3, 4, 5\}$ for $\pi_0 = \frac{1}{3}$ (e. g. the triangle test, table 3) and $\pi_0 = \frac{1}{2}$ (e. g. the duo-trio test, table 4), respectively. All the same, different effect sizes are considered. We extract some of the data and add some more values to give those combinations of n and k that result in a power of 90%, 95% and 99%, respectively. Schlich (1993) defines the effect size to be the relative success probability above chance, i. e. an effect size of, e. g., 50% would state that the mean success probability of the assessors is given by $\pi_0 + (1 - \pi_0) * 0.5$, while an effect size of 25% is given by $\pi_0 + (1 - \pi_0) * 0.25$. Table 1 gives the corresponding values for the case of $\pi_0 = \frac{1}{3}$, while table 2 gives these values for $\pi_0 = \frac{1}{2}$. The same effect sizes as in Schlich (1993) and Brockhoff (2002) are considered here. Note that the respective significance level in all tables is $\alpha = 5\%$.

	$1-\beta$		90%			95%			99%	
	effect size	25%	37.5%	50%	25%	37.5%	50%	25%	37.5%	50%
k=1	n	74/81	35/39	18/22	95/102	43/47	23/27	130/135	58/60	34
	nk	74/81	35/39	18/22	95/102	43/47	23/27	130/135	58/60	34
k=2	n	45	20	13	53/57	26/28	15	76/79	37	21
	nk	90	40	26	106/114	52/56	30	152/158	74	42
k=3	n	34	15	9	39/43	20	13	58	28	16/18
	nk	102	45	27	117/129	60	39	174	84	48/54
k=4	n	27	12	8	34	17	10	47/51	25	15
	nk	108	48	32	136	68	40	188/204	100	60
k=5	n	23	11	7	31	16	9	45	23	13
	nk	115	55	35	155	80	45	225	115	65

Table 1: Numbers n and nk for given k resulting in a power of at least $1-\beta$ for different effect sizes for a success probability by pure guessing of $\pi_0 = \frac{1}{3}$.

	$1-\beta$	90%			95%			99%		
	effect size	25%	37.5%	50%	25%	37.5%	50%	25%	37.5%	50%
k=1	n	132/143	60/65	30/35	169/175	76/78	42	228/239	102/104	56/58
	nk	132/143	60/65	30/35	169/175	76/78	42	228/239	102/104	56/58
k=2	n	74	35	19	93	41/45	22/25	128	59	34
	nk	148	70	38	186	82/90	44/50	256	118	68
k=3	n	54	26	14	67/69	31	17	94/98	42/44	24/26
	nk	162	78	42	201/207	93	51	282/294	126/132	72/78
k=4	n	44	20	11	54	26	15	76/79	37	20/22
	nk	176	80	44	216	104	60	304/316	148	80/88
k=5	n	35/38	17	10	46	23	12	66	31	19
	nk	175/190	85	50	230	115	60	330	155	95

Table 2: Numbers n and nk for given k resulting in a power of at least $1-\beta$ for different effect sizes for a success probability by pure guessing of $\pi_0 = \frac{1}{2}$.

Assume that we want to design a triangle test and are that we are interested in finding a large effect, i. e. an effect of 50%, with probability 95%. Stressing table 1 we may either use 23 assessors without replications (the value 27 will be explained in the next paragraph) or 15 assessors with two replications each, giving 30 assessments altogether, or 13 assessors with three replications each (39 assessments), or ten assessors with four (40), or finally 9 assessors with 5 (45). From these different combinations, the investigator might choose the one which results in lowest costs. These values are valid for maximal heterogeneity of the assessors as mentioned before. However, note that these values are only valid for this large effect size – for smaller effect sizes, the number of assessments is much larger, as it can be seen from the tables. Furthermore, these tables only address the power of the test. This means that we may rely on the fact that we might have found similarity in case we cannot reject the null hypothesis. As well, we could rely on the fact that the products are different in case we could reject the null hypothesis. However, we could not necessarily rely on the estimated effect size. We might have induced a larger variability to the estimator, such that the outcomes are less reliable. This will be addressed later in this paper in more details.

For the interpretation of tables 1 and 2, first of all it has to be noticed that the power in Brockhoff's (2002) paper is not monotonically increasing with an increasing number of assessors. Even though it usually should be, for the binomial test this is not the case. This is due to the discreteness of the binomial test distribution which leads to a waste in the significance level α. More precise, this means that in case we perform a level-α test, in almost every case the true probability to mistakenly reject the null hypothesis is actually smaller than α. Further details can be found in any textbook giving an introduction to statistics and in particular non-parametric statistics. In our context, this means that the power of the test might indeed decrease with an increasing number of assessors. In tables 1 and 2, wherever appropriate we give the smallest number of n and nk, respectively, for which the power does not fall below $1 - \beta$, as well as the numbers for which we never will fall again below this value while increasing n. For further details we refer to the tables provided by Brockhoff (2002).

Of course even smaller effect sizes might be of interest, recalling that an effect size of 25% means that the assessors will perceive the difference in every fourth trial. However, for an effect size of 12.5%, say, the number get very large and most likely far beyond any reasonable number of assessments in practice. Even for $k = 1$, for the triangle test we would need 283/301, 356/372 and 495/511 assessors to obtain a power of 90%, 95% and 99%, respectively, while these number are even larger for the duo-trio test, namely 546/565, 674/695 and 938/948, respectively. Knowing that reducing the number of assessors while increasing the number of replications will even increase the total number of assessments, this seems not feasible at all. Numbers of this size are also derived by Schlich (1993) who states, e. g., that for an effect size of 10% in the triangle test 447 assessors are needed to ensure a power of 90%.

One important interpretation of these tables is that we might indeed achieve reasonable power even for small numbers of assessors. For instance in table 1, given an effect size of 50%, 13 assessors replicating twice will result in a power of 90%, while only 8 additional assessors are needed to increase this power to 99%. In the latter case, we would need 34 assessors if no replications are considered to gain the same power, i. e. we save 13 assessors by adding 8 additional assessments only. However, note that this holds only for an effect size of 50%! An effect of this size could mean that all assessors will actually perceive the difference in every second trial. The other extreme case for this effect size is that half of the assessors will never find any difference and will therefore only guess, while the other half will always perceive it. Of course, many other assessor performances may result in the same effect size. Using these values to design the experiment, accepting the null hypothesis and therefore claiming similarity implicitly means that the products are still considered being similar whenever assessors will detect a difference every second time! There might be applications aiming at similarity in which this might indeed be appropriate, while we point out that we are mostly interested in small effect sizes like 25%, say, or even less! A 25% effect means that the difference is on average found in one out of four trials, which might already be a proportion to worry about.

Comparing the results of the two tables, it can be found that the increase of total assessments needed for an identical power while increasing the number of replications is smaller in the $\pi_0 = \frac{1}{2}$ case in table 2 than for the one with $\pi_0 = \frac{1}{3}$ in table 1. This is due to the fact that these values depend on the heterogeneity of the assessors, which is much more restricted in the duo-trio test than in the triangle test. If the assessors were absolutely homogeneous, meaning that they all had an identical success probability, we would not lose any information by considering replications instead of different assessors. However, here we consider the worst case in which the assessors are as heterogeneous as possible, meaning that one part of the assessors has success probability π_0 while the other has 1. Hence for the duo-trio-test these probabilities are $\frac{1}{2}$ respectively 1, such that they differ from each other by $\frac{1}{2}$ in probability, while they do by $\frac{2}{3}$ for the triangle test. Due to the experiment only, assessors performing in a duo-trio test will therefore usually be more homogeneous than those in a triangle test will – the success probabilities may obviously vary more.

Finally it has to be stated that these results are valid only if the sampling of the assessors has been appropriate. It has to be assured that the assessors indeed represent the overall population they are meant to represent. This means that a fully randomized sampling from this overall population would be most favorable. Since unfortunately this seems to be unfeasible in practice, even more concerns have to be given that the sampling is nevertheless reasonably done. While sampling details are beyond the scope of this paper, we want to point out that in case of replications, the outcomes of the test will be heavily influenced by an inappropriate sampling whenever this does not represent the population to a large extent. In this case, we would not recommend the use of replications at all in order to account for an absence of assessors. Nevertheless, whenever possible replications will be reasonable if they can be done in addition to the initial design.

Confidence intervals

The results of the former sections can also be transformed into results for confidence intervals. However, in this case we might be less interested in the total number of successes itself, but rather in the relativ number of successes given the considered number of assessments. This theoretical overall value will be denoted by π, say, while the most commonly used estimator for this value is given by

$$\hat{\pi} = \frac{X}{nk},$$

which is once again a random variable. From the results of the former section, we easily find that

$$\mathbf{E}(\hat{\pi}) \quad = \quad \mathbf{E}(P_1) \tag{4}$$

and

$$\mathbf{Var}(\hat{\pi}) \quad = \quad \frac{k-1}{nk}\mathbf{Var}(P_1) + \frac{1}{nk}\mathbf{E}(P_1)(1 - \mathbf{E}(P_1)), \tag{5}$$

from which again the advantage of an as small as possible number of replications with respect to the variance can be found. Contrariwise, the expectation still remains unaffected. Furthermore, for a fixed n it can directly be seen that

$$\mathbf{Var}(\hat{\pi}) \quad = \quad \frac{1}{n}\mathbf{E}(P_1)(1 - \mathbf{E}(P_1))$$

for $k = 1$, while for $k \to \infty$ we find

$$\mathbf{Var}(\hat{\pi}) \quad \longrightarrow \quad \frac{1}{n}\mathbf{Var}(P_1), \tag{6}$$

so that $\frac{1}{n}\mathbf{Var}(P_1)$ is a lower limit of the variability for any given n.

Let us now consider a given number n where $k = 1$. We are interested in different combinations of the number of assessors, denoted by m for a moment, and replications giving a variance not larger than the setting with n and $k = 1$. The question arises to what extent the number of assessors may decrease while k tends to infinity. From the formulae above we find that due to this restriction, we end up with the condition

$$\frac{1}{m}\mathbf{Var}(P_1) \quad \leq \quad \frac{1}{n}\mathbf{E}(P_1)(1 - \mathbf{E}(P_1))$$

which is equivalent to

$$m \quad \geq \quad n\frac{\mathbf{Var}(P_1)}{\mathbf{E}(P_1)(1 - \mathbf{E}(P_1))}. \tag{7}$$

If the distribution of P_1 is known, this value can explicitly be determined as it will be shown later in this paper.

As it has been mentioned before and is shown in the appendix, for a given number of assessors n the variability of $\hat{\pi}$ monotonically decreases with an increasing number of replications k in all cases of practical interest. Hence an additional replication for the assessor should be considered whenever this is possible.

Having said that, we should as well note that an increase of replications is only reasonable to some extent. From (6)we find that the variance will not fall below a certain lower limit, and whenever we already have a reasonable number of replications, adding another one might not result in a noticeable decrease of the variance anymore. As well, we see that consistency can only be achieved by letting the number of assessors n tend to infinity. In that case, we have

$$\mathbf{Var}(\hat{\pi}) \quad \longrightarrow \quad 0 \qquad (n \to \infty).$$

Knowing from (4) that $\hat{\pi}$ is an unbiased estimator for the expectation of P_1, we have proven the consistency of this estimator for $n \to \infty$. Note that this does not depend at all on the value of k. However, as it has been said before, in practice we will always have to confine ourselves to a finite and usually rather small n, and therefore some replications might decrease the variance to a reasonable extent.

Constructing confidence limits using, e. g., the Tschebycheff inequality, a 95% confidence interval for π is given by $\hat{\pi} \pm 2\sqrt{\mathbf{Var}(\hat{\pi})}$, while a 99% interval is given by $\hat{\pi} \pm 3\sqrt{\mathbf{Var}(\hat{\pi})}$. Hence, the confidence intervals get larger as the variance of π increases and therefore, in order to minimize the size of these intervals, for a fixed number of assessments nk the number of replications k is to be minimized. Note that in practice we do not know the variance of $\hat{\pi}$ but have to estimate it. In this case, the intervals given above become $\hat{\pi} \pm 2\sqrt{\hat{\mathbf{Var}}(\hat{\pi})}$ and $\hat{\pi} \pm 3\sqrt{\hat{\mathbf{Var}}(\hat{\pi})}$.

Note that usually better confidence intervals exist than those derived by means of the Tschebycheff inequality. This is mainly due to the Tschebycheff inequality being very conservative, such that the intervals can be improved in case of more restrictive assumptions, e. g. regarding the distribution of P_1. For the difference test we have at least the knowledge about P_1 being a distribution on the interval $[\pi_0, 1]$ only. Nevertheless every reasonable interval will depend on the variance of the respective estimator and therefore the general conclusion drawn here still holds for any other (reasonable) case.

Two special cases

In this section, we take two different models into account, which have been frequently considered with respect to replicated difference tests, see also Brochhoff (2002). One of these has previously been referred to as the mixture binomial model, the other one as the beta-binomial model. We will determine the variance of $\hat{\pi}$ in these cases, depending on n and k, such that we may compare the variabilities induced from different strategies in designing the experiment.

The mixture binomial model

The first distribution considered here is derived from the assumption that a proportion δ, say, of the assessors will detect the difference with a common success probability π_1, say, where $\pi_1 > \pi_0$. The other proportion of the assessors will only be guessing, such that they have a success probability of π_0. More formally, we may write this as

$$\begin{aligned} P_i &= \begin{cases} \pi_0 & \text{with probability } 1-\delta, \\ \pi_1 & \text{with probability } \delta, \pi_1 > \pi_0 \end{cases} \\ &= (1-\Delta)\,\pi_0 + \Delta\pi_1 \end{aligned}$$

where $\mathbf{L}(\Delta) = \mathbb{B}(1, \delta)$. From the latter formulation, the reason for this model being referred to as a mixture binomial model is obvious.

In this case, it can be easily shown that

$$\mathbf{E}(P_i) \;=\; \pi_0 + (\pi_1 - \pi_0)\delta,$$

depending on π_1 as well as on δ, such that different combinations of these two values may result in an identical mean and therefore in the same effect size. This may equivalently be written as

$$\mathbf{E}(P_i) - \pi_0 \;=\; (\pi_1 - \pi_0)\delta \tag{8}$$

and

$$\delta \;=\; \frac{\mathbf{E}(P_i) - \pi_0}{\pi_1 - \pi_0}, \tag{9}$$

respectively. As stated in Brockhoff (2002), the variance of P_i is given by

$$\begin{aligned}\mathbf{Var}(P_i) &= \mathbf{E}(P_i^2) - (\mathbf{E}(P_i))^2 \\ &= \pi_0^2 + \delta\left(\pi_1^2 - \pi_0^2\right) - (\pi_0(1-\delta) + \pi_1\delta)^2 \\ &= (1-\delta)\,\delta(\pi_1 - \pi_0)^2.\end{aligned}$$

Using (8), this becomes

$$\mathbf{Var}(P_i) \;=\; \left(\frac{1}{\delta} - 1\right)(\mathbf{E}(P_i) - \pi_0)^2 .$$

For a fixed effect size respectively mean of P_i, recalling $\delta \in [0, 1]$ we find that the variability increases with a decrease of δ and the other way round decreases with its increase. Hence it is obvious that the variability is minimized by $\delta = 1$ and with it $\pi_1 = \mathbf{E}(P_i)$. This means that all assessors have exactly the same success probability. Thus the heterogeneity is minimal, all assessments are independent from each other, no matter whether they stem from replications or not.

In contrast, from (9) we find that the variability is maximized by minimizing $\frac{\mathbf{E}(P_i)-\pi_0}{\pi_1-\pi_0}$ respectively by maximizing π_1. Since $\pi_1 \in [\mathbf{E}(P_i), 1]$, for a fixed effect size we find that the variability is maximized in case of $\pi_1 = 1$ and $\delta = \frac{\mathbf{E}(P_i)-\pi_0}{1-\pi_0}$. This means that for this worst case scenario, δ is identical with the effect size, which is actually defined by this fraction (Schlich, 1993). All the same, here the assessor heterogeneity is maximal: The success probabilities of the assessors are as far apart from each other as possible. One group of relative size δ has the maximal success probability of 1 while the other group of relative size $1 - \delta$ has the minimal success probability of pure guessing, π_0.

For practical applications, this means that we have to be aware of a large variability whenever we expect the assessors to be quite different, i. e. some of them being (almost) pure guessers while the others are (almost) sure perceivers. In such a case, an increase of the number of assessments might be particularly worthwhile in order to decrease this variance and therefore get more reliable results.

Now using the results from (1) and (2), for the mean and variance of X we get

$$\mathbf{E}(X) = nk\Big(\pi_0 + (\pi_1 - \pi_0)\delta\Big)$$

and

$$\begin{aligned}\mathbf{Var}(X) &= n(k^2 - k)\Big((1-\delta)\,\delta(\pi_1 - \pi_0)^2\Big) \\ &\quad + nk\Big(\pi_0 + (\pi_1 - \pi_0)\delta\Big)\Big(1 - \pi_0 - (\pi_1 - \pi_0)\delta\Big),\end{aligned}$$

respectively, as well as we get for $\hat{\pi}$

$$\mathbf{E}(\hat{\pi}) = \pi_0 + (\pi_1 - \pi_0)\delta$$

and

$$\begin{aligned}\mathbf{Var}(\hat{\pi}) &= \frac{k-1}{nk}\Big((1-\delta)\,\delta(\pi_1 - \pi_0)^2\Big) \\ &\quad + \frac{1}{nk}\Big(\pi_0 + (\pi_1 - \pi_0)\delta\Big)\Big(1 - \pi_0 - (\pi_1 - \pi_0)\delta\Big). \end{aligned} \tag{10}$$

This equation will be used later to calculate different combinations of n and k leading to identical variabilities and therefore to identical confidence intervals.

Here, using (7) results in

$$m \geq \frac{(1-\delta)\,\delta(\pi_1 - \pi_0)^2}{(\pi_0 + (\pi_1 - \pi_0)\delta)(1 - (\pi_0 + (\pi_1 - \pi_0)\delta))}\, n. \tag{11}$$

Considering the worst case scenario in terms of maximal assessor heterogeneity for the triangle test, it was shown above that we have $\pi_0 = \frac{1}{3}$ and $\pi_1 = 1$. In this case, the effect size is identical with the value of δ. Hence the lower limits for the number of assessors needed even if k tends to infinity are easily found to be $\frac{n}{3}$, $\frac{3}{7}n$ and $\frac{n}{2}$ for an effect size of 25%, 37.5% and 50%, respectively. To say it the other way round: No matter how many replications we consider, for an effect size of, e. g., 50% we will never be able to achieve the same small variance compared to the respective unreplicated design with less than $\frac{n}{2}$ assessors.

The corrected Beta-binomial model

The other distribution considered for P_i is the corrected Beta-distribution. The Beta-distribution is frequently used due to its nice properties, meaning that most calculations of interest end with not too complicated terms. However, the Beta-distribution itself is not applicable within this context, since it is a distribution on the interval $[0,1]$ instead of $[\pi_0,1]$. However, Brockhoff (2002) proposes an adaption of the probability density such that values smaller than π_0 will not occur, while the distribution remains easy to handle. The Beta-distribution has two parameters, $\alpha > 0$ and $\beta > 0$, say. Brockhoff (2002) gives details about how to estimate these parameters from the data.

The corrected Beta-binomial model can be expressed such that the random variable P_i is derived by means of

$$P_i = \pi_0 + (1-\pi_0)Q_i,$$

where $\mathbf{L}(Q_i) = \text{Beta}(\alpha,\beta)$ for $i = 1,2,3,...$ Hence, knowing that

$$\mathbf{E}(Q_i) \quad = \quad \frac{\alpha}{\alpha+\beta}$$

and

$$\mathbf{Var}(Q_i) \quad = \quad \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)},$$

for the expectation and variance of P_i we directly get

$$\mathbf{E}(P_i) \quad = \quad \pi_0 + (1-\pi_0)\frac{\alpha}{\alpha+\beta}$$

and

$$\mathbf{Var}(P_i) \quad = \quad (1-\pi_0)^2\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)},$$

respectively. Again, note that the same effect size may stem from different combinations of the parameters α and β.

Using (1) and (2) once more, we find the expectation and variance of the total number of successes to be

$$\mathbf{E}(X) = nk\left(\pi_0 + (1-\pi_0)\frac{\alpha}{\alpha+\beta}\right)$$

and

$$\begin{aligned}\mathbf{Var}(X) &= n(k^2-k)\left((1-\pi_0)^2\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}\right)\\ &\quad +nk\left(\pi_0 + (1-\pi_0)\frac{\alpha}{\alpha+\beta}\right)\left((1-\pi_0)\left(1-\frac{\alpha}{\alpha+\beta}\right)\right)\end{aligned}$$

as well as we get for $\hat{\pi}$ that

$$\mathbf{E}(\hat{\pi}) = \pi_0 + (1-\pi_0)\frac{\alpha}{\alpha+\beta} \tag{12}$$

and

$$\begin{aligned}\mathbf{Var}(\hat{\pi}) &= \frac{k-1}{nk}\left((1-\pi_0)^2\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}\right)\\ &\quad +\frac{1}{nk}\left(\pi_0 + (1-\pi_0)\frac{\alpha}{\alpha+\beta}\right)\left((1-\pi_0)\left(1-\frac{\alpha}{\alpha+\beta}\right)\right).\end{aligned} \tag{13}$$

The equations developed here will be used in what follows to compare different strategies in designing the experiment with respect to the choice of n and k.

How to design an experiment?

From the results given before, two main strategies are directly derived for the design of an experiment investigating in differences between products:

- Use as many assessments as possible.
- For a given number of assessments nk, let the number of assessors n be as large as possible and restrain the number of replications k to a minimum necessary to achieve the intended number of trials.

Besides these main recommendations, however, we are now considering some kind of mixed strategy. In doing so, we assume that the use of replications instead of additional assessors might be much cheaper and convenient to the investigator. Therefore it might occur that, e. g., we have to decide whether we have $n = 30$ and $k = 1$ or rather $n = 20$ and $k = 2$. Now, the total number of assessments should no longer be considered as fixed, such that we have to trade off between the different possibilities we have. If we knew the true underlying distribution of P_i, we could easily compare the variances of, e. g., $\hat{\pi}$ and choose the option with the smallest value under the given circumstances. Unfortunately, in practice we do not even know the class of distributions that includes the one of the P_i, let alone the exact parameters. Therefore, we give the corresponding values for different parameter settings within the two models considered here. No general recommendations might be drawn from these, but in practical circumstances the investigator might rely on the following tables to consider the worst-case scenario, or maybe she / he is willing to append some further assumptions with respect to the distribution, maybe based on some prior knowledge regarding the products and their differences.

First of all, using (10) we consider the worst case scenario for the mixture binomial model within the triangle test, which means that we have maximal heterogeneity between the assessors, i. e. $\pi_1 = 1$ as it has been shown earlier. This is also referred to as the common limit model (Brockhoff, 2002), since it is the asymptotic limit model for the three possible models considered within his paper, when the assessor heterogeneity tends to its maximal value. These models are namely the mixture binomial, the Beta-binomial and the generalized linear mixed model which is not considered within this paper. In this case, it has been shown earlier as well that the effect size is equal to the value of δ. Considering the triangle test, we finally have $\pi_0 = \frac{1}{3}$, such that we can easily determine the variability of the estimator for the effect size depending on different true effect sizes according to (10). Starting with a particular n and $k = 1$, we determine the essential number of assessors for $k = 2, 3, ..., 10$ resulting at least in an as small variance as the first case. The results for an effect size of 50%, 37.5% and 25% including the respective total numbers of assessments are given in tables 3, 4 and 5, respectively.

k	1	2	3	4	5	6	7	8	9	10
n	10	8	7	7	6	6	6	6	6	6
nk	10	16	21	28	30	36	42	48	54	60
n	15	12	10	10	9	9	9	9	9	9
nk	15	24	30	40	45	54	63	72	81	90
n	20	15	14	13	12	12	12	12	12	11
nk	20	30	42	52	60	72	84	96	108	110
n	25	19	17	16	15	15	15	15	14	14
nk	25	38	51	64	75	90	105	120	126	140
n	30	23	20	19	18	18	18	17	17	17
nk	30	46	60	76	90	108	126	136	153	170
n	35	27	24	22	21	21	21	20	20	20
nk	35	54	72	88	105	126	147	160	180	200
n	40	30	27	25	24	24	23	23	23	22
nk	40	60	81	100	120	144	161	184	207	220
n	45	34	30	29	27	27	26	26	25	25
nk	45	68	90	116	135	162	182	208	225	250
n	50	38	34	32	30	30	29	29	28	28
nk	50	76	102	128	150	180	203	232	252	280

Table 3: Combinations of n and k resulting in a variance not larger than the respective case with $k = 1$ for the worst case scenario of the triangle test with an effect size of 50%.

k	1	2	3	4	5	6	7	8	9	10
n	10	8	7	6	6	6	6	5	5	5
nk	10	16	21	24	30	36	42	40	45	50
n	15	11	10	9	9	8	8	8	8	8
nk	15	22	30	36	45	48	56	64	72	80
n	20	15	13	12	11	11	11	10	10	10
nk	20	30	39	48	55	66	77	80	90	100
n	25	18	16	15	14	14	13	13	13	13
nk	25	36	48	60	70	84	91	104	117	130
n	30	22	19	18	17	16	16	15	15	15
nk	30	44	57	72	85	96	112	120	135	150
n	35	26	22	21	20	19	18	18	18	18
nk	35	52	66	84	100	114	126	144	162	180
n	40	29	25	23	22	21	21	20	20	20
nk	40	58	75	92	110	126	147	160	180	200
n	45	33	28	26	25	24	23	23	23	22
nk	45	66	84	104	125	144	161	184	207	220
n	50	36	31	29	28	27	26	26	25	25
nk	50	72	93	116	140	162	182	208	225	250

Table 4: Combinations of n and k resulting in a variance not larger than the respective case with $k = 1$ for the worst case scenario of the triangle test with an effect size of 37.5%.

k	1	2	3	4	5	6	7	8	9	10
n	10	7	6	5	5	5	5	5	5	5
nk	10	14	18	20	25	30	35	40	45	50
n	15	11	9	8	8	7	7	7	7	6
nk	15	22	27	32	40	42	49	56	63	60
n	20	14	12	10	10	9	9	9	9	9
nk	20	28	36	40	50	54	63	72	81	90
n	25	17	14	13	12	12	11	11	11	11
nk	25	34	42	52	60	72	77	88	99	110
n	30	21	17	16	15	14	13	13	13	12
nk	30	42	51	64	75	84	91	104	117	120
n	35	24	20	18	17	16	16	15	15	15
nk	35	48	60	72	85	96	112	120	135	150
n	40	27	23	20	19	18	18	17	17	17
nk	40	54	69	80	95	108	126	136	153	170
n	45	30	26	23	21	21	20	19	19	19
nk	45	60	78	92	105	126	140	152	171	190
n	50	34	28	26	24	23	22	21	21	21
nk	50	68	84	104	120	138	154	168	189	210

Table 5: Combinations of n and k resulting in a variance not larger than the respective case with $k = 1$ for the worst case scenario of the triangle test with an effect size of 25%.

From tables 3-5 we find that indeed a reasonable decrease in the number of assessors can be obtained by means of replications. However, it is easily seen that introducing replications results in a very strong increase in the total number of assessments. For an effect size of 50% and $n = 50$, say, we would have to carry out more than 50% of additional tests in order to shorten the essential number of assessors at less than 25% ($k = 2$ in table 3). To reduce the number of assessors at 40% to a total of 30, a 200% increase of the number of assessments to 150 is needed ($k = 5$ in table 3). Note that the minimal number of assessors cannot fall below $\frac{n}{2}$ in this case as has been shown earlier. Similar conclusions can be drawn from tables 4 and 5 for an effect size of 37.5% and 25%, respectively. Note that in these cases the minimal number of assessors will not fall below $\frac{3n}{7}$ and $\frac{n}{3}$, respectively.

In tables 6-8, the respective values for the two-alternative-forced-choice difference tests like, e. g., the duo-trio test can be found. Here, $\pi_0 = \frac{1}{2}$ holds.

k	1	2	3	4	5	6	7	8	9	10
n	10	7	6	5	5	5	5	5	5	4
nk	10	14	18	20	25	30	35	40	45	40
n	15	10	9	8	7	7	7	7	7	6
nk	15	20	27	32	35	42	49	56	63	60
n	20	14	12	10	10	9	9	9	9	8
nk	20	28	36	40	50	54	63	72	81	80
n	25	17	14	13	12	12	11	11	11	10
nk	25	34	42	52	60	72	77	88	99	100
n	30	20	17	15	14	14	13	13	13	12
nk	30	40	51	60	70	84	91	104	117	120
n	35	24	20	18	17	16	15	15	15	14
nk	35	48	60	72	85	96	105	120	135	140
n	40	27	23	20	19	18	18	17	17	16
nk	40	54	69	80	95	108	126	136	153	160
n	45	30	25	23	21	20	20	19	19	18
nk	45	60	75	92	105	120	140	152	171	180
n	50	34	28	25	24	23	22	21	21	20
nk	50	68	84	100	120	138	154	168	189	200

Table 6: Combinations of n and k resulting in a variance not larger than the respective case with $k = 1$ for the worst case scenario of the duo-trio test with an effect size of 50%.

k	1	2	3	4	5	6	7	8	9	10
n	10	7	6	5	5	4	4	4	4	4
nk	10	14	18	20	25	24	28	32	36	40
n	15	10	8	7	7	6	6	6	6	6
nk	15	20	24	28	35	36	42	48	54	60
n	20	13	11	10	9	8	8	8	8	7
nk	20	26	33	40	45	48	56	64	72	70
n	25	16	13	12	11	10	10	10	9	9
nk	25	32	39	48	55	60	70	80	81	90
n	30	20	16	14	13	12	12	11	11	11
nk	30	40	48	56	65	72	84	88	99	110
n	35	23	19	16	15	14	14	13	13	13
nk	35	46	57	64	75	84	98	104	117	130
n	40	26	21	19	17	16	16	15	15	14
nk	40	52	63	76	85	96	112	120	135	140
n	45	29	24	21	19	18	17	17	16	16
nk	45	58	72	84	95	108	119	136	144	160
n	50	32	26	23	21	20	19	19	18	18
nk	50	64	78	92	105	120	133	152	162	180

Table 7: Combinations of n and k resulting in a variance not larger than the respective case with $k = 1$ for the worst case scenario of the duo-trio test with an effect size of 37.5%.

k	1	2	3	4	5	6	7	8	9	10
n	10	6	5	4	4	4	4	3	3	3
nk	10	12	15	16	20	24	28	24	27	30
n	15	9	7	6	6	5	5	5	5	5
nk	15	18	21	24	30	30	35	40	45	50
n	20	12	10	8	8	7	7	6	6	6
nk	20	24	30	32	40	42	49	48	54	60
n	25	15	12	10	9	9	8	8	8	8
nk	25	30	36	40	45	54	56	64	72	80
n	30	18	14	12	11	10	10	9	9	9
nk	30	36	42	48	55	60	70	72	81	90
n	35	22	17	14	13	12	11	11	11	10
nk	35	44	51	56	65	72	77	88	99	100
n	40	24	19	16	15	14	13	12	12	12
nk	40	48	57	64	75	84	91	96	108	120
n	45	28	22	19	17	15	15	14	14	13
nk	45	56	66	76	85	90	105	112	126	130
n	50	30	24	20	18	17	16	15	15	15
nk	50	60	72	80	90	102	112	120	135	150

Table 8: Combinations of n and k resulting in a variance not larger than the respective case with $k = 1$ for the worst case scenario of the duo-trio test with an effect size of 25%.

Stressing (11) once again, we find the lower limit for the number of necessary assessors to be $\frac{n}{3}$, $\frac{3n}{11}$ and $\frac{n}{5}$ for an effect size of 50%, 37.5% and 25%, respectively. In comparison with tables 3-5, note that the two different groups of assessors are not as different from each other in the two-alternative-forced-choice case as they are in the three-alternative-forced-choice case considered before, since $\pi_1 - \pi_0$ is now equal to $\frac{1}{2}$ instead of $\frac{2}{3}$ as before. Therefore the heterogeneity is not as large anymore and the loss in terms of increasing variance is not as large as it was before in the three-alternative-forced-choice cases. Nevertheless, considering an effect size of 50% once more, we still have to pay with 36% additional assessments to reduce the number of assessors at 32% only ($k = 2$ in table 6). In order to halve the number of assessors, we would have to double the number of assessments ($k = 4$ in table 6). Again, similar results can be found for other effect sizes as shown in tables 7 and 8.

The considerations given here hold only for the worst case scenario described above, i. e. the common limit model for maximal assessor heterogeneity. Assuming the mixture binomial model, it can be easily found from (10) that in case of $\delta = 1$, the variability does not depend on the number of replications, but on the number of total assessments nk only. Hence it does not matter at all whether we consider replications or not. For any number of replications, the total number of assessments needed to obtain the same small variability remains the same. For the corrected Beta-binomial model, this holds only in a certain limit, namely if both α and β tend to zero as it can be seen from (13). However, besides being the other extreme cases for the two special cases considered here, these ones are not of much use for applications, since we will never assume all assessors to have identical success probabilities. The truth usually will be somewhere in between these extreme cases, while we will never now what the correct model will be. Hence for the design of the experiment, it seems reasonable to consider the worst case given above. To give an impression about the behavior of other circumstances, we consider now some different circumstances for the three-alternative-forced-choice difference tests, all of which result in the same effect size of 50%.

To start with, we reconsider the mixture binomial model. The common limit model stressed before is equal to the mixture binomial model with $\pi_1 = 1$ and $\delta = 0.5$, cf. table 3. The case of no heterogeneity is given by $\delta = 1$ and $\pi_1 = \frac{2}{3}$. Recalling that hence $\mathbf{E}(P_i) = \frac{2}{3}$ and stressing (9), we find that the same effect size is all the same given for the settings $\pi_1 = \frac{9}{10}$ and $\delta = \frac{10}{17}$ respectively $\pi_1 = \frac{3}{4}$ and $\delta = \frac{4}{5}$. Tables 9 and 10 give the same values as considered before for these combinations.

k	1	2	3	4	5	6	7	8	9	10
n	10	7	6	6	5	5	5	5	5	5
nk	10	14	18	24	25	30	35	40	45	50
n	15	11	9	8	8	7	7	7	7	7
nk	15	22	27	32	40	42	49	56	63	70
n	20	14	12	11	10	10	9	9	9	9
nk	20	28	36	44	50	60	63	72	81	90
n	25	17	15	13	12	12	12	11	11	11
nk	25	34	45	52	60	72	84	88	99	110
n	30	21	17	16	15	14	14	13	13	13
nk	30	42	51	64	75	84	98	104	117	130
n	35	24	20	18	17	17	16	16	15	15
nk	35	48	60	72	85	102	112	128	135	150
n	40	27	23	21	20	19	18	18	17	17
nk	40	54	69	84	100	114	126	144	153	170
n	45	31	26	24	22	21	20	20	19	19
nk	45	62	78	96	110	126	140	160	171	190
n	50	34	29	26	24	23	23	22	22	21
nk	50	68	87	104	120	138	161	176	198	210

Table 9: Combinations of n and k resulting in a variance not larger than the respective case with $k = 1$ for the mixture binomial model in the triangle test and an effect size of 50% derived from the setting $\pi_1 = \frac{9}{10}$ and $\delta = \frac{10}{17}$.

k	1	2	3	4	5	6	7	8	9	10
n	10	6	5	4	3	3	3	3	3	3
nk	10	12	15	16	15	18	21	24	27	30
n	15	9	7	6	5	5	4	4	4	4
nk	15	18	21	24	25	30	28	32	36	40
n	20	12	9	7	6	6	5	5	5	5
nk	20	24	27	28	30	36	35	40	45	50
n	25	15	11	9	8	7	7	6	6	6
nk	25	30	33	36	40	42	49	48	54	60
n	30	17	13	11	9	9	8	8	7	7
nk	30	34	39	44	45	54	56	64	63	70
n	35	20	15	13	11	10	9	9	8	8
nk	35	40	45	52	55	60	63	72	72	80
n	40	23	17	14	12	11	10	10	9	9
nk	40	46	51	56	60	66	70	80	81	90
n	45	26	19	16	14	13	12	11	10	10
nk	45	52	57	64	70	78	84	88	90	100
n	50	29	21	18	15	14	13	12	12	11
nk	50	58	63	72	75	84	91	96	108	110

Table 10: Combinations of n and k resulting in a variance not larger than the respective case with $k = 1$ for the mixture binomial model in the triangle test and an effect size of 50% derived from the setting $\pi_1 = \frac{3}{4}$ and $\delta = \frac{4}{5}$.

As was to be expected, from tables 9 and 10 it can be seen that the increase of the number of assessments nk in order to reduce the number of assessors gets smaller the smaller π_1 and the larger δ. The results in table 9 are quite similar to those in table 3, while in table 10 we are close to the situation of no heterogeneity and therefore an reasonable reduction of the number of assessors can be achieved by a relative minor increase of assessments. Nevertheless, we state once again that we will never be sure in applications that this situation is truly given, therefore an use of these tables in order to determine the number of assessor and replications might be quite risky.

Finally, we also consider two circumstances within the Beta-binomial model. It has been mentioned before that the case of extreme heterogeneity is given by α and β tending to zero. As pointed out in Brockhoff (2002), a measure of heterogeneity is given by

$$\theta = \frac{1}{\alpha + \beta + 1},$$

while the heterogeneity increases with an increase of θ. Hence, the assessors are most homogeneous if α and β tend to infinity. From (12) we easily find that the effect size is given by $\frac{\alpha}{\alpha+\beta}$, such that we get $\alpha = \beta$ in order to get an effect size of 50%. Bi and Ennis (1998) use values of $\theta = 0, 0.1, ..., 0.5$, from which we chose 0.5 and 0.2, resulting in $\alpha = \beta = \frac{1}{2}$ and $\alpha = \beta = 2$, respectively. The results are quite similar to those found in tables 9 and 10, respectively, such that we do not give them here for brevity. In fact, even though this does not hold for these particular values, we may derive identical tables for the mixture binomial and the Beta-binomial model. The variance of $\hat{\pi}$ in (10) and (13), respectively, is decomposed in two addends. In each case, the latter one depends on the total number of assessments nk and the mean of P_i only. Since the effect size has been fixed to 50% here, this latter term is therefore identical for both models. The former addend depends in both models on the number of replications relative to the total number of assessments given by $\frac{k-1}{nk}$. For given δ and π_1, say, we can find α and β such that the variance of $\hat{\pi}$ is identical. For this purpose, we would have to solve

$$(1-\delta)\delta(\pi_1 - \pi_0)^2 = (1-\pi_0)^2 \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}.$$

Recalling that in our case $\alpha = \beta$, this simplifies to

$$(1-\delta)\delta(\pi_1 - \pi_0)^2 = (1-\pi_0)^2 \frac{\alpha^2}{4\alpha^2(2\alpha+1)},$$

resulting in a simple quadratic equation after inserting the appropriate values of π_0, π_1 and δ. Even though for the effect size of 50%, this is particularly simple, we will find corresponding values for any possible combination. The other way round, we might as well fix α and β and determine the corresponding values for the mixed binomial model.

This result similarly holds in more general. From (5) we find that we have the same structure for any distribution of P_1, the latter term depending on the effect size while the former one depends on the variability. This means that whenever we may vary the parameters of the distribution such that variability and mean can be independently adjusted, we will be able to equate this model with any other one. To say it the other way round: It does not matter which model is considered, we may always find a setting of the other reasonable models which leads to the same variability of the estimator for π. This nicely supports the findings of Brockhoff (2002), who stated as well that the outcomes with respect to different models do not differ too much. However, he was rather stressing the power of the corresponding test, while we consider the variability of the estimators.

Discussion

The use of replications in difference tests is sometimes considered whenever the number of available assessors seems to be too small and cannot be increased due to, e. g., budget restrictions. Doubts have arisen that the application of commonly used statistical procedures is still reasonable in this case, cf. Brockhoff and Schlich (1998). Though Kunert and Meyners (1999) showed that the binomial test with nk observations will not violate the significance level α of the test whenever the experiment is properly designed, we have shown that the power of the test will increase whenever the number of assessments is large enough such that the power is larger than 50%. On the other hand, the power will decrease in case that either the number is too small or the effect size is smaller than assumed and hence the true power of the test is smaller than 50%. Hence adding replications after having the experiment designed such that it would work out without these replications will always be a plus for the analysis and the reliability of the interpretation.

Unfortunately, chances are this will not be considered very often in practice.

Interests are rather given to a reduction of the number of assessments in all, while often preferred in assessors. From our results it has to be stated that both the estimation of the effect size and the prove of similarity will be disadvantageously influenced by adding replications and reducing the number of assessors while the number of assessments remains the same. This holds to a larger extent the more heterogeneous the assessors are. We have developed different formulae to determine the variability of reasonable estimators. Furthermore, we have compared the power of different designs according to Brockhoff's (2002) approach. For both cases, we have determined different combinations of n and k that lead to the same variability respectively the same power. From these it can always be found that a decrease in the number of assessors has to be paid for by an increase of the total number of assessments nk. However, it might be reasonable to use these results according to the respective costs in hiring assessors and performing additional assessments. Since the results for the different criteria do not coincide, an a-priori decision is essential about which information will be of main interest. If both are, the approach with more assessors and, in case, less replications should be chosen.

Using replications, serious concerns should be given to the sampling of the assessors. The results presented in this paper only hold if the sampling is appropriate with respect to the overall population that should be represented. Most favorable would be a full randomized sampling from the population of interest. Whenever this is not possible, representativeness of the sample will be questionable. Even though this is a problem for non-replicated difference tests as well, this holds to a much larger extent whenever we use replications. In the latter case, we rely much more heavily on single assessors – if those few assessors have been poorly chosen (due to chance or an inappropriate procedure), the results will be heavily influenced by even letting these assessors replicate. Thus we might end up with serious misinterpretations. Hence, whenever the sampling of the assessors might be questionable, we do not recommend to rely on replications at all. Even if the outcomes of a non-replicated design using the same sampling might be doubtful as well, chances are that we will nevertheless get more reliable results.

Assuming that the sampling has been properly done such that no more concerns have to be given to this point of the experimental design, we will now discuss the design of an experiment as well as the interpretation of the outcomes. In a triangle test, suppose we are interested in an effect size of 37.5% (or larger) and want to restrict the type II error rate β to 10% respectively guarantee a power of at least 90%. From table 1 we find that without replications we need 35 assessors to achieve these values. In order to reduce the number of assessors, we might use as well 20 assessors only but letting them perform the test twice, such that we end up with 40 assessments. All the same, we might consider 15, 12 and 11 assessors only with 3, 4 and 5 replications and 45, 48 and 55 total assessments, respectively. From these combinations, the investigator might choose that one that fits his requirements best – given he is only interested in showing differences or similarity! If she/he is interested in estimating the true effect size in case differences have been shown, table 4 has to be stressed. From this we find that, according to the effect size of interest, the variance of the estimator of interest derived from 20 assessors performing the triangle test twice only accounts of about 27 / 28 assessments of different assessors. For the other combinations, with 3 replications this value is given by about 24, while for $k = 4, 5$ it is 20 only. Hence, using table 1 to design the experiment will result in an increased uncertainty about this estimator.

On the other hand, we might still want to use replications. Table 4 shows as well that the variability derived from 35 assessments from different assessors is the same as from 26 assessors with 2 replications as from 22, 21 and 20 assessors with 3, 4 and 5 replications. Hence, the total number of assessments has to be increased to 52, 66, 84 and 100, respectively. Now again, the investigator might choose one of these combinations according to his interest, and we claim that this will indeed result in the same certainty of the estimator. In addition, since these values are larger than those derived from table 1, the true power of the test will increase! In this case, the use of replications might indeed be useful, while the total number of assessment definitely has to be reasonably increased. In application, 35 assessors doing the test once only might be cheaper and better feasible than letting 20 assessors replication 5 times. Nevertheless, if the latter case is feasible, this might be a plus since the true power rises over 95% as it can be seen from table 1. In contrast, not increasing the total number of assessments while using replications will definitely reduce the power as well as the reliability of any estimators of interest.

The outcomes of such a series of triangle tests might be very different. First of all, it might appear that we cannot prove differences and therefore claim similarity. In case the experiment has been designed according to table 1, we can claim this similarity with a maximal error rate of β. If the total number of assessments has not been increased but replications have been used, this does not hold any longer, since the power of the test will decrease as it can been seen from our tables 1 and 2 as well as from those provided by Brockhoff (2002). and we usually cannot claim similarity since the power of the test is unknown. Note that, considering an effect size of 37.5%, we explicitly agree to claim similarity at a 5%-level in case of not being able to prove differences with 16 assessors only, but 5 replications each such that we have 80 assessments. Still, concerns have to be given to the sampling once again!

Second, we might find differences. As Kunert and Meyners (1999) have shown, this is the easiest case since these differences have been proven to a significance level α whenever the experiment has been properly designed. In this case, it does not matter whether or not the assessments stem from replications.

Finally, having proven differences, we might want to estimate the effect size. In this case, the experiment should have been designed according to, e. g., table 4.

Then the variability of the estimator of interest is identical no matter which of the corresponding combinations of n and k has been chosen, meaning that a reduction of assessors goes together with a reasonable number of assessments. On the other hand, if we have not appropriately increased the number of assessments, we will end up with a much less reliable estimator. This means that the variability of this estimator is quite large and therefore any reasonable confidence interval will be as well. Frankly spoken, estimating the effect size to be, e. g., 37.5% might mean that the effect size could be 5% or 70% as well. Hence the uncertainty is very large. The use of more different assessors and less replications or an appropriate number or additional assessments to respect for the replications would have notably decreased the uncertainty. Then, with the same estimated effect size of 37.5%, we might be sure that the true one is not smaller than 25% and not larger than 50%, say.

As a final remark, we want to point that using replications more powerful tests exits besides the binomial test. These tests base on the likelihood-ratio approach and are much more complicated to carry out. Due to this, the binomial test will be the mostly used one in applications, such that details about other tests are beyond the scope of this paper.

To summarize, we state that together with an appropriate sampling, replications might be used to reduce the number of assessors to some extent. However, to derive similarly reliable results, the total number of assessments has to be reasonably increased. In that case, these combinations of n and k might even increase the power of the test. On the other hand, if the number of assessments cannot be heavily increased, we propose to use non-replicated tests whenever this is possible. Furthermore, main concerns have to be given to an appropriate sampling of the assessors. Even though this holds for non-replicated tests as well, it holds for replicated designs to a larger extent.

Acknowledgements

This paper was written while the first author visited the Department of Mathematics and Physics at the Royal Veterinary and Agricultural University (KVL) in Frederiksberg. He gratefully acknowledges the support of the Rudolf-Chaudoire foundation for this visit, the hospitality of the people at KVL and the general financial support of the Deutsche Forschungsgemeinschaft (SFB 475, "Reduction of complexity for multivariate data structures") of his work.

References

Bi, J. and Ennis, D.M. (1999a) *Beta-binomial tables for replicated difference and preference tests.* Journal of Sensory Studies 14, 347-368.

Bi, J. and Ennis, D.M. (1999b) *The power of sensory discrimination methods used in replicated difference and preference tests.* Journal of Sensory Studies 14, 289-302.

Brockhoff, P.B. (2002) *The statistical power of replications in difference tests.* submitted to Food Quality and Preference.

Brockhoff, P.B. and Schlich, P. (1998) *Handling replications in discrimination tests.* Food Quality and Preference 9, 303-312.

Ennis, D.M. and Bi, J. (1998) *The Beta-binomial model: accounting for inter-trial variation in replicated difference and preference tests.* Journal of Sensory Studies 13, 389-412.

Hunter, E.A., Piggott, J.R. and Lee, M.K.Y. (2000) *Analysis of discrimination tests.* Agro-industrie et methodes statistiques, Pau, January 19-21, 2000.

Kunert, J. and Meyners, M. (1999) *On the triangle test with replications.* Food Quality and Preference 9, 303-312.

McCullagh, P. and Nelder, J.A. (1989) *Generalized linear models.* 2nd ed., London, Chapman and Hall.

Meyners, M. (2002) *On the number of perceivers in a triangle test with replications.* In: C. Duby and J.P. Cassar (eds.): Actes des 7èmes Journées Européennes Agro-Industrie et Méthodes Statistiques, Lille, January 16-18, 2002, 85-89.

Schlich, P. (1993) *Risk tables for discrimination tests.* Food Quality and Preference 4, 141-151.

Appendix

Proposition: For a fixed number of assessors n, $\mathbf{Var}(\hat{\pi})$ monotonically decreases with an increasing number of replications k, while the monotonicity is strict whenever P_1 is not almost sure equal to 1.

Proof: Due to the distribution of P_1 being restricted on the interval $[\pi_0, 1]$ where $\pi_0 > 0$, $0 < \mathbf{E}(P_1) \leq 1$. From $0 < P_1 \leq 1$ it follows $P_1^2 \leq P_1$ and hence $\mathbf{E}(P_1^2) \leq \mathbf{E}(P_1)$. Equality holds if and only if $P_1 = 1$ almost sure.

Now let n be fixed and $0 < k < l$ be two possible numbers of replications. $\hat{\pi}_k$ and $\hat{\pi}_l$ denote the estimate $\hat{\pi}$ given k and l replications, respectively. Then we have

$$
\begin{aligned}
&\mathbf{Var}(\hat{\pi}_k) - \mathbf{Var}(\hat{\pi}_l) \\
&= \frac{1}{n}\left[\left(\frac{k-1}{k} - \frac{l-1}{l}\right)\mathbf{Var}(P_1) + \left(\frac{1}{k} - \frac{1}{l}\right)\mathbf{E}(P_1)(1-\mathbf{E}(P_1))\right] \\
&= \frac{1}{n}\left[\frac{kl - l - lk + k}{kl}\mathbf{Var}(P_1) + \frac{l-k}{kl}\mathbf{E}(P_1)(1-\mathbf{E}(P_1))\right] \\
&= \frac{l-k}{nkl}\left[\mathbf{E}(P_1)(1-\mathbf{E}(P_1)) - \mathbf{Var}(P_1)\right] \\
&= \frac{l-k}{nkl}\left[\mathbf{E}(P_1) - (\mathbf{E}(P_1))^2 - \mathbf{Var}(P_1)\right] \\
&\geq \frac{l-k}{nkl}\left[\mathbf{E}(P_1^2) - (\mathbf{E}(P_1))^2 - \mathbf{Var}(P_1)\right] \qquad (14) \\
&= \frac{l-k}{nkl}\left[\mathbf{Var}(P_1) - \mathbf{Var}(P_1)\right] \\
&= 0
\end{aligned}
$$

Equality in (14) holds if and only if $\mathbf{E}(P_1) = \mathbf{E}(P_1^2)$, i. e. if and only if P_1 is almost sure equal to 1. This completes the proof of theorem 1.

Note that a similar result under slightly more restrictive conditions is also given by, e. g., McCullagh and Nelder (1989). Also note that P_1 being almost sure 1 means that each assessor will always succeed, i. e. the product differences must be extremely large. This case might not be of much importance, nevertheless it is already intuitively clear that the variability of X respectively $\hat{\pi}$ does not depend on the number of replications considered, since it will always be zero.