Published by Oxford University Press on behalf of the International Epidemiological Association
International Journal of Epidemiology 2010;**39**:60–68
doi:10.1093/ije/dyn366

THEME: METHODOLOGY, THEORY AND APPLICATION

LQAS: User Beware

Dale A Rhoda,[1]* Soledad A Fernandez,[1] David J Fitch[2] and Stanley Lemeshow[1]

Accepted 10 December 2008

Background Researchers around the world are using Lot Quality Assurance Sampling (LQAS) techniques to assess public health parameters and evaluate program outcomes. In this paper, we report that there are actually two methods being called LQAS in the world today, and that one of them is badly flawed.

Methods This paper reviews fundamental LQAS design principles, and compares and contrasts the two LQAS methods. We raise four concerns with the simply-written, freely-downloadable training materials associated with the second method.

Results The first method is founded on sound statistical principles and is carefully designed to protect the vulnerable populations that it studies. The language used in the training materials for the second method is simple, but not at all clear, so the second method sounds very much like the first. On close inspection, however, the second method is found to promote study designs that are biased in favor of finding programmatic or intervention success, and therefore biased against the interests of the population being studied.

Conclusion We outline several recommendations, and issue a call for a new high standard of clarity and face validity for those who design, conduct, and report LQAS studies.

Keywords Lot quality assurance sampling, quality assurance, healthcare, sampling studies, evaluation studies, intervention studies, prevalence, immunization

Background

In a recent review, Robertson and Valadez reported that Lot Quality Assurance Sampling (LQAS) techniques were used in more than 800 health-related surveys between 1984 and 2004, mostly in developing countries.[1] LQAS is supposed to provide a rapid and inexpensive estimate of the prevalence of a specific condition such as a malady or a successful intervention. The topics investigated in the studies described in Robertson and Valadez[1] were as diverse as immunization coverage, post-disaster public health, neonatal tetanus mortality and service delivery quality management.

In this article, we report that there are actually 2 methods being called LQAS in the world today, and that one of them is badly flawed. The first method, which we review and endorse, is founded on sound statistical principles and is carefully designed to protect the vulnerable populations that it studies. It poses a null hypothesis that the malady is widespread or that the intervention has not been successful, and only rejects that null in the face of strong evidence.[2,3] In recent years, the first method has been overshadowed by a second approach, which sounds very much like the first, but reverses the role of the null and alternative hypotheses. Rather than protect the population at risk, it poses a null hypothesis that the population is healthy or that an intervention has been successful, and then accepts the null unless there is overwhelming evidence

[1] College of Public Health, The Ohio State University, Columbus, OH, USA.

[2] Instituto de Investigaciones Económicas y Sociales, Universidad Rafael Landívar, Guatemala.

* Corresponding Author. 330 Blandford Drive, Worthington, OH 43085, USA. E-mail: rhoda.4@osu.edu

to reject it. Accepting a null hypothesis is always a statistical error. Simply put, the second method is biased toward concluding that interventions have reached their goals before they actually do.

According to Robertson and Valadez,[1] there were fewer than 50 LQAS surveys being reported per year before 1999, but the number climbed to more than 200 surveys in 2004. They suggest that one factor in the expanded use of LQAS is the 'availability of practical manuals and guidelines' and they cite 'difficult-to-understand statistical explanations that were not helpful to public health professionals interested in field applications' as one early 'impediment in applying the method'.[1] The second method is taught using some 'practical manuals and guidelines' that are freely available on the Internet. Besides reversing the traditional direction of LQAS hypothesis tests, the manuals may lead trainees to believe that small sample techniques are much more powerful than they actually are. They report very low error rates based on a complicated and unstated definition of 'error' rather than the simple definition that trainees are likely to infer from the over-simplified materials.

While we applaud the work that has gone into developing practical manuals and the effort to make them available at low cost, we are alarmed that some important LQAS principles have been lost along the way. We fear that faulty LQAS conclusions may be used to deny interventions or preventative services to people who desperately need them. This article compares and contrasts the two LQAS methods and concludes with recommendations for those who design, carry out and report LQAS studies.

An overview of LQAS

Health ministries and international development organizations are often interested in estimating the prevalence of certain conditions or characteristics. These might include:

- prevalence of a disease or health condition,
- proportion of the target population that has received an intervention,
- proportion of the population that knows a risk-related fact (e.g. AIDS can be transmitted through sexual contact),
- proportion of mothers trained to properly mix oral rehydration solution.

In this article, we use the example of estimating the proportion of the population who have received a particular vaccination. The health ministry may wish to accomplish the following two goals.

(1) Estimate the overall population proportion vaccinated for an entire region.
(2) Identify smaller districts within the region that have especially high or especially low proportions. Those with low proportions of vaccination may require special interventions. Those with high proportions, might not need special intervention any longer. Furthermore, those with high proportions might serve as models of 'best practices'.

If the inquiring agency were able to allocate unlimited resources to the task of evaluation, then both goals could be met using a census or using large sample surveys in each district. Where resources are limited, however, it is not always possible to obtain precise estimates of both the regional proportion and the individual district proportions.

The LQAS solution to this problem is to perform small sample studies in each district and then aggregate the results to estimate the regional proportion. LQAS studies use sample sizes on the order of dozens per district rather than hundreds, so the confidence interval for each district proportion is very large. When the estimates from multiple districts are pooled, the straightforward formula for the estimate of population proportion from a stratified sample yields a precise regional estimate from imprecise district estimates.

Although confidence intervals for individual districts are not especially informative, the study organizers often wish to identify the districts whose proportion exceeds a particular threshold. At the district level, LQAS may be understood as a straightforward application of a binomial hypothesis test.

We believe that the process of designing an LQAS study should include the following.

(1) Select P_0, the proportion threshold of interest and construct the null hypothesis. It is traditional to assume that the population is not healthy, or not being served adequately and to only reject that assumption in the face of strong evidence to the contrary. In the vaccination example, the null hypothesis is that the proportion of persons vaccinated in the district, $P_d \leqslant P_0$.
(2) Select an acceptable upper bound for the probability of type I error (α). A type I error would occur if the investigator concluded that $P_d > P_0$ when, in fact, it is not.
(3) Select P_2, a second proportion threshold, for the purposes of specifying either the power of the test ($1-\beta$), or the probability of type II error (β). A type II error occurs anytime the investigator fails to reject the null hypothesis when, in fact, $P_d > P_0$. Select an acceptable value of β for the probability of type II error (β) if the true district proportion $P_d > P_2$.
(4) Use an LQAS table (e.g. from Lemeshow and Taber[3]) to determine which combinations of sample size (n) and decision threshold (d^*) will provide tests that meet the type I and type II constraints for P_0 and P_2.
(5) Randomly sample n individuals from each district. If at least d^* of the sampled individuals have been served, then the investigator has strong evidence to conclude that $P_d > P_0$.

Otherwise, the investigator fails to reject the null hypothesis that $P_d \leqslant P_0$.

(6) Combine the counts from individual districts to compute an aggregate prevalence and confidence interval for the entire region. If the figures from individual districts vary widely, the average prevalence may not be very meaningful. In that case, it might be helpful to report the range of figures from the districts.

For the purpose of clarity in this article, we assume that higher proportions indicate intervention success, as would be true with the prevalence of vaccination. If the issue at hand is prevalence of a malady rather than an intervention, then of course lower proportions will indicate intervention success. In that case, we can conceive of a test where higher proportions are good news by estimating the proportion of persons who do not have the malady rather than the proportion who do.

Example

Suppose a health administrator wishes to estimate the proportion of the region that has received a particular vaccination and to identify those districts where she can be confident that $P_d > 50\%$. She might set P_0 to be 50%, and $\alpha = 10\%$. The null hypothesis is that $P_d \leqslant 50\%$. She might choose to control the type II error rate such that $\beta = 10\%$ at $P_2 = 80\%$. The rejection criterion will have less than 10% probability of failing to reject the null hypothesis when $P_d > 80\%$ and <10% probability of rejecting the null when $P_d \leqslant 50\%$. The values $n = 19$ and $d^* = 13$ satisfy these criteria. If 13 or more vaccinated persons are found in a district's sample, then the administrator rejects the null hypothesis and concludes confidently that $P_d > 50\%$. Otherwise, she fails to reject the null hypothesis. (Note that the exact 90% lower confidence bound for a proportion given 13 successes in 19 trials falls above 50%, whereas the lower 90% confidence bound given 12 successes in 19 trials falls below 50%.)

Features of LQAS study designs

Several features of LQAS study designs warrant careful attention.

LQAS designs may be summarized with operating characteristic curves

Figure 1 shows the operating characteristic curve of the $n = 19$, $d^* = 13$ LQAS design. The abscissa represents P_d, the true proportion of vaccinated persons in the district. The height of the curve indicates the probability of obtaining 13 or more successes in 19 independent trials at each value of P_d. When $P_d = P_0$, the curve attains a height of α, the maximum probability of rejecting the null hypothesis if it is true. Note that $\alpha < 10\%$ at $P_d = 50\%$ for this curve. Any value of P_d that is $>P_0$ could be selected for P_2. At those points, the height of the curve represents $1-\beta$, or the power of the study design to reject the null hypothesis. Note that when $P_d = 80\%$, $1-\beta > 90\%$ which means that $\beta < 10\%$, so this design meets the criteria listed above.

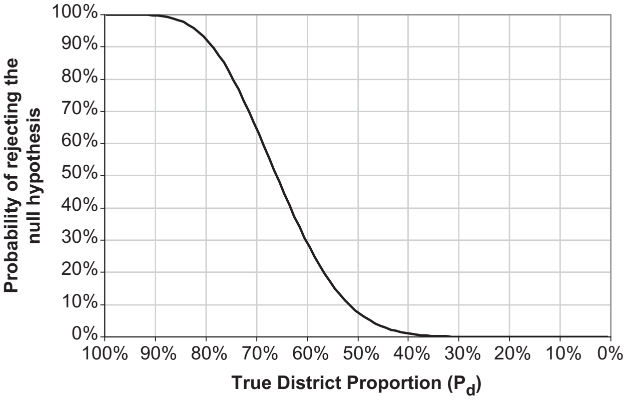

Figure 1 Operating characteristic curve for the $n = 19$, $d^* = 13$ study design using the LQAS method that protects vulnerable populations

Because of the discrete nature of the binomial distribution, only discrete values of α and β will be possible with LQAS designs. When n or d^* changes, the achievable values of α and β change discretely. In order to achieve $\alpha = \beta = 10\%$ exactly at P_0 and P_2, a very large value for n would be necessary. It is customary to choose values for α and β and then select combinations of n and d^* that provide error rates no larger than, but sometimes smaller than, α at P_0 and β at P_2.

LQAS designs can, and should, protect vulnerable populations

In some situations, administrators may use LQAS results to allocate resources. They might devote extra resources to districts that do not show evidence of having reached P_0, and they might shift resources away from districts that appear to have crossed that threshold. This makes the direction of the null hypothesis very important.

To see why this is so, consider the implications of type I and type II errors for the population under study. When the null assumes that $P_d \leqslant P_0$, type I errors mean that resources are mistakenly withdrawn from districts that have not yet reached P_0. Fortunately, we set a low value for α, so the administrator will rarely withdraw resources from needy districts. Type II errors occur when P_d is above P_0 and the administrator continues to devote resources to districts that have already reached P_0. This may be an inefficient use of resources, but it does not endanger the population as clearly as a type I error does. We see in Figure 1 that the decision rule has low power for rejecting the null hypothesis when P_d is between 50% and 80%, so we expect type II errors

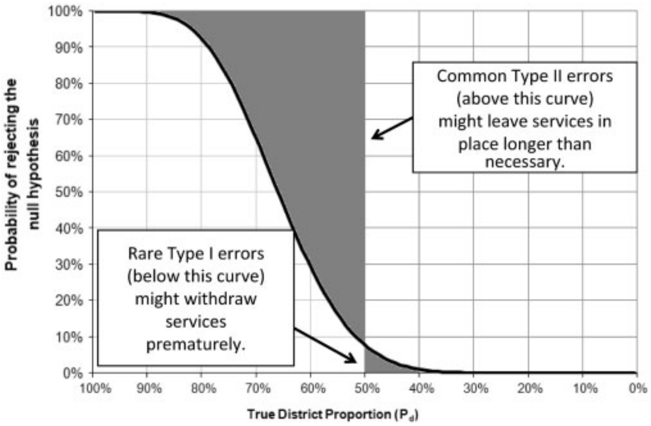

Figure 2 With a well-constructed null hypothesis, type II errors are preferable to type I errors

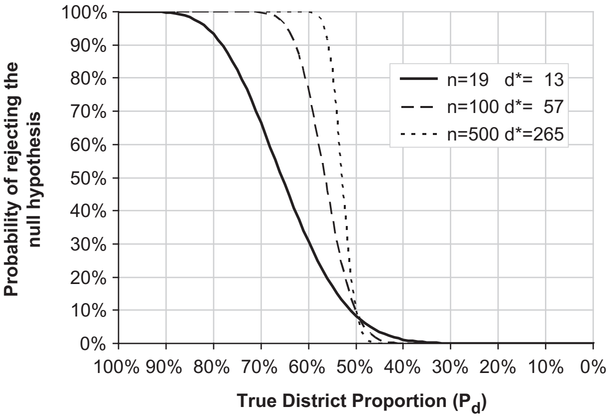

Figure 3 Three LQAS designs where $P_0 = 50\%$ and $\alpha = 10\%$

to be common when P_d is in that range. Indeed the probability of type II error is as high as $1-\alpha$ when P_d is just above P_0. Figure 2 indicates that the vulnerable population will prefer common errors of type II to rare errors of type I.

On the other hand, if the null hypothesis states that the district is being adequately served, ($P_d \geqslant P_0$), then the decision rule will require strong evidence to conclude otherwise. The sample proportion will need to be quite a bit smaller than P_0 to conclude that $P_d < P_0$. Rare type I errors will devote extra resources to districts that do not need them, and common type II errors will withdraw resources from needy districts. This design is biased against the persons being studied and biased in favour of finding that the intervention has reached its goals. We feel strongly that reversing the null hypothesis in this way is a disservice to the population at risk. In many cases, the people being studied with LQAS are living in poverty. Economic, political and environmental circumstances may be stacked against their odds of living a healthy life. We feel strongly that the LQAS study design should not be stacked against them, too.

Therefore, the null hypothesis should be constructed to assume that the people are not healthy or not well served, and the study design should require strong evidence to conclude otherwise. P_0 should be selected carefully, and a small value should be chosen for α. Note that in our example, $(d^*)/n = 13/19 = 0.684$ or 68.4%. This design is conservative in that it assumes that the proportion of persons who have been vaccinated is $\leqslant$50% and it only rejects that null hypothesis if 68.4% or more of the persons sampled have been vaccinated. The design requires strong evidence to conclude that the vaccination programme has reached the threshold of 50%.

When using small values of *n*, LQAS designs have low power

Recall that the height of the operating characteristic curve represents the probability of rejecting the null hypothesis. When P_d is larger than, but near P_0, the $n = 19$, $d^* = 13$ rule has very low power. When $P_d = 60\%$, the power is only 30%, so there is a 70% chance of making a type II error. When $P_d = 70\%$, there is a 33% chance of making a type II error. One method of addressing this problem is to choose an LQAS design with larger values for n and d^*. Figure 3 shows the operating characteristic curves for three designs where $P_0 = 50\%$ and $\alpha = 10\%$. Higher values of n and d^* result in designs that are more powerful for rejecting the null hypothesis at values of $P_d > P_0$. Another way to address the problem is to use a double-sampling design that surveys additional persons if the sample proportion from the first n individuals is too close to P_0 to draw a confident conclusion.[3]

Some problems evident in LQAS training materials

In 2002, Valadez and Devkota published a table entitled 'Optimal LQAS Decision Rules for Sample Sizes of 12–30 and Coverage Benchmarks or Average Coverage of 20%–95%'.[4] That table is reproduced here as Table 1. The table has subsequently been used to train numerous people in LQAS techniques. It appeared in Valadez *et al.*[5] with small differences in the footnote and title wording. More recently, it appeared in training materials that have been made freely available on the Internet.[6–8]

Bearing in mind the features of LQAS designs that were articulated above, we have several grave concerns with Table 1 and with the LQAS designs and training materials that are based upon it.

Concern 1: The null hypotheses in Table 1 are biased against vulnerable populations

This is our most serious concern. The table and its associated training materials avoid statistical jargon so they never state a null hypothesis, per se, but we can infer what the null must be by looking at the thresholds in the top row of the table and the sample

Table 1 LQAS Table from Valadez *et al.*[8]

LQAS table: decision rules for sample sizes of 12–30 and coverage targets/average at 10%–95%

	Average coverage (baselines)/annual coverage target (monitoring and evaluation)																	
Sample Size	10%	15%	20%	25%	30%	35%	40%	45%	50%	55%	60%	65%	70%	75%	80%	85%	90%	95%
12	NA	NA	1	1	2	2	3	4	5	5	6	7	7	8	8	9	10	11
13	NA	NA	1	1	2	3	3	4	5	6	6	7	8	8	9	10	11	11
14	NA	NA	1	1	2	3	4	4	5	6	7	8	8	9	10	11	11	12
15	NA	NA	1	2	2	3	4	5	6	6	7	8	9	10	10	11	12	13
16	NA	NA	1	2	2	3	4	5	6	7	8	9	9	10	11	12	13	14
17	NA	NA	1	2	2	3	4	5	6	7	8	9	10	11	12	13	14	15
18	NA	NA	1	2	2	3	5	6	7	8	9	10	11	11	12	13	14	16
19	NA	NA	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
20	NA	NA	1	2	3	4	5	6	7	8	9	11	12	13	14	15	16	17
21	NA	NA	1	2	3	4	5	6	8	9	10	11	12	13	14	16	17	18
22	NA	NA	1	2	3	4	5	7	8	9	10	12	13	14	15	16	18	19
23	NA	NA	1	2	3	4	6	7	8	10	11	12	13	14	16	17	18	20
24	NA	NA	1	2	3	4	6	7	9	10	11	13	14	15	16	18	19	21
25	NA	1	2	2	4	5	6	8	9	10	12	13	14	16	17	18	20	21
26	NA	1	2	3	4	5	6	8	9	11	12	14	15	16	18	19	21	22
27	NA	1	2	3	4	5	7	8	10	11	13	14	15	17	18	20	21	23
28	NA	1	2	3	4	5	7	8	10	12	13	15	16	18	19	21	22	24
29	NA	1	2	3	4	5	7	9	10	12	13	15	17	18	20	21	23	25
30	NA	1	2	3	4	5	7	9	11	12	14	16	17	19	20	22	24	26

NA: not applicable, meaning LQAS cannot be used in this assessment because the coverage is either too low or too high to assess a supervision area.[8]
Notes: Lightly shaded cells indicate where α or β errors are ≤ 10%. Darker cells indicate where α or β errors are ≤ 15%.[8]

We do not recommend using this table for LQAS study design. Except for minor wording changes, this is the same as Table 1 in Valadez and Devkota.[4] Sample size (n) is listed in the leftmost column. Values of d^* are listed in the table. Prevalence thresholds or 'coverage targets' are listed in the top row. Note that $(d^*)/n <$ threshold for every entry in the table. These decision rules implicitly assume that the population proportion exceeds the coverage benchmark and only conclude otherwise if the sample mean is dramatically below the threshold. They result in study designs that are biased toward concluding that an intervention has been successful.

proportions that are represented by $(d^*)/n$. Note that the design where $n = 19$ and $d^* = 13$ appears in Table 1 and it is associated with a threshold of 80%, not the 50% that we used above. According to the instructions that accompany Table 1 in Valadez *et al.*[4], if 13 or more out of 19 have been vaccinated, then the investigator should conclude that the population proportion is at least 80%. Recall that 13/19 = 68.4%. This study design clearly establishes the null hypothesis to be that the proportion is ⩾80%, and requires strong evidence (a sample proportion below 68.4%) to conclude otherwise.

For every sample size and threshold in Table 1, the proportion represented by $(d^*)/n$ is smaller than the threshold being tested, so all of the study designs assume that the proportion exceeds the threshold and only conclude otherwise if the sample proportion is much lower than the threshold. These designs set the bar too low! They are biased to conclude that the intervention programme has been successful and will only conclude otherwise in the face of strong evidence.

The null hypothesis for the $n = 19$, $d^* = 13$ rule in Table 1 is that $P_d \geqslant 80\%$ and the alternative hypothesis is that $P_d < 80\%$. Figure 4 shows the operating characteristic curve for this design. Figure 5 shows that, in this case, a type I error is made when the administrator erroneously concludes that $P_d < 80\%$, and continues to devote resources to the district even though it has reached the 80% threshold. The much more common type II error fails to reject the null

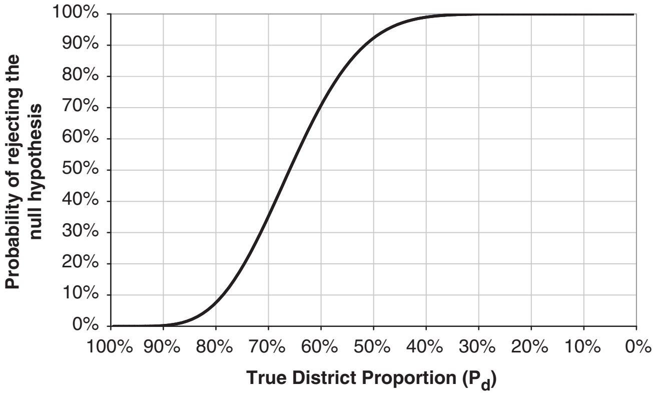

Figure 4 Operating characteristic curve for the $n = 19$, $d^* = 13$ rule using the LQAS method that reverses the roles of the null and alternative hypotheses

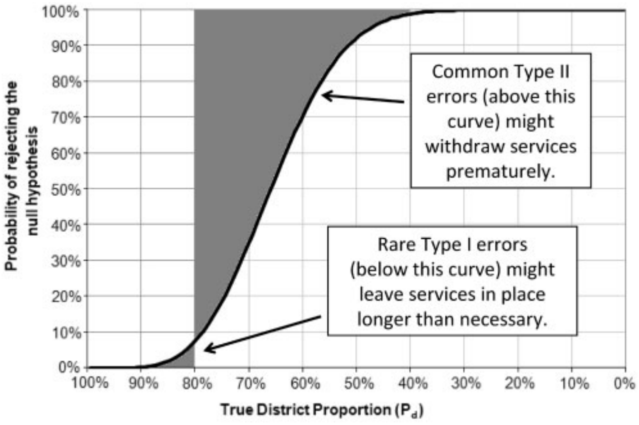

Figure 5 When the null hypothesis is biased toward the intervention, type II errors are biased against the vulnerable population

hypothesis and concludes that the district has reached 80% prevalence, when, in fact, $P_d < 80\%$. Having erroneously concluded that the goals have been met, the health administrator might withdraw resources from this district when the population is still struggling to meet the 80% goal. Table 2 summarizes the implications of this concern both in general, and for the vaccination example.

Reversing the null hypothesis in this manner goes against longstanding tradition in quality assurance sampling. The use of LQAS in public health is modeled on the work of Dodge and Romig in the manufacturing domain.[9,10] In manufacturing, batches or 'lots' of identical parts are supplied by a part 'producer'. Each lot should be inspected by the 'consumer' of the parts to verify that the quality of the lot is acceptable. If at least d^* out of n sampled parts are within specification, then the lot is accepted by the consumer. If not, then one of several consequences follows: either the lot is rejected outright and sent back to the producer, or in some cases, every piece in the lot is inspected before being used.

The first sentence of the introduction in the first edition of Dodge and Romig's book says 'It has long been recognized, where sampling instead of complete inspection is used, that certain errors or risks are unavoidable.'[9] They use the terms 'consumer's risk' and 'producer's risk' to indicate that, by inspecting only a sample rather than every part in every lot, the consumer assumes some risk that they will accept a 'bad' lot and the producer assumes some risk of having good lots rejected. In the language of public health, by evaluating a sample of persons rather than evaluating every eligible individual, the public assumes some (consumer's) risk that the intervention will be declared a success prematurely, and resources will be withdrawn. Likewise, the health ministry assumes some (producer's) risk that resources will unnecessarily continue to be expended in a region where the programme's goals have already been met.

Dodge and Romig make it clear that the first priority of their inspection method is to protect the consumer. 'The first requirement for the method will therefore be in the form of a definite assurance against passing any unsatisfactory lot that is submitted for inspection. [...] For the first requirement, there must be specified at the outset a value for the tolerance per cent defective as well as a limit to the chance of accepting any submitted lot of unsatisfactory quality. The latter has, for convenience, been termed the Consumer's Risk...'.[9] Although the Table 1 study designs look superficially like Dodge and Romig designs, they differ fundamentally from those designs in that they put the first priority on limiting the producer's risk, rather than that of the vulnerable public.[11]

Concern 2: The study designs in these training materials purport to have low error rates (this assertion is very likely to be misinterpreted)

The LQAS training materials based on Table 1 claim that their study designs have low error rates, with both α and β less than 10% in many cases. We are concerned because the materials provided to the trainees do not clearly define what they mean by 'error'. Without a clear definition, we feel that it is likely that the trainees will adopt a simple and logical definition of 'error':

- intuitive type I error: concluding that the district has reached the threshold, when it has not, or
- intuitive type II error: concluding that the district has not reached the threshold, when it has.

Instead, the definition of 'error' that results in α and β below 10% is more complicated, and it includes both P_0 and P_2. For the $n = 19$, $d^* = 13$ study in the LQAS training materials, the definition of error is something like as given below.

- Conclude that the district has reached the threshold of 80% when, in fact, the true proportion lies below 50%.
- Conclude that the district has not reached the threshold of 80% when in fact the true proportion-lies above 80%.
- If the true proportion lies between 50% and 80%, then any conclusion is possible and none are regarded as 'errors'.

We feel that this language is misleading for several reasons.

Concern 2a: Table 1 only lists one threshold, the upper threshold, where the designs control the probability of type I error

Because the reader or trainee is not informed about the other threshold, at which the design controls the

Table 2 Comparison of the two approaches to LQAS study designs

	LQAS method with protective null hypothesis (We advocate this type of design)		**Method based on Table 1 (We do not advocate this type of design)**	
	In general	Vaccination example	In general	Vaccination example
Null hypothesis	$P_d \leqslant P_0$	$P_d \leqslant 50\%$ Vaccination prevalence is low	$P_d \geqslant P_0$	$P_d \geqslant 80\%$ Vaccination prevalence is high
Alternative hypothesis	$P_d > P_0$	Vaccination prevalence is high	$P_d < P_0$	Vaccination prevalence is low
Conclusion if more than d^* individuals with the condition of interest are found in the random sample of size n	$P_d > P_0$ Reject the null hypothesis and declare intervention success at the $P_d = P_0$ level	Conclude that the intervention has been successful for at least 50% of the district	$P_d \geqslant P_0$ Accept the null hypothesis & declare intervention success at the $P_d = P_0$ level	Conclude that the intervention has been successful for at least 80% of the district
Conclusion if fewer than d^* out of n individuals have the condition of interest	Fail to reject the null hypothesis and continue intervention efforts	There is not strong enough evidence to conclude that at least 50% of the district has been vaccinated	$P_d < P_0$ Reject the null hypothesis and continue intervention efforts	There is strong evidence that less than 80% of the district has been vaccinated
Consequence of a (rare) type I error This will happen with probability $\leqslant \alpha$	Declare intervention success prematurely	Possibly withdraw resources prematurely	Fail to recognize intervention success in a timely manner	Leave intervention resources in place longer than necessary to reach the 80% goal
Consequences of a (common) type II error. The probability of type II error depends on P_d. It can be as high as $1-\alpha$ when P_d is near P_0	Fail to recognize intervention success in a timely manner	Leave intervention resources in place longer than necessary to reach the 50% goal	Declare intervention success prematurely	Possibly withdraw resources prematurely

In one method, the population is protected by the null hypothesis and by the relatively common type II errors. In the other method, a common type II error might declare intervention success and withdraw resources prematurely. We advocate the method on the left side of the table.

probability of type II error, it would be impossible for the trainee to infer the correct definition of 'error'. To be fair, we must point out that the correct definition of 'error' does appear in the appendix to the trainer's manual.[12] But, the complete and correct definition of 'error' does not appear in the material provided to the trainee.

Concern 2b: Trainees are not told anything about the possibility of making an error when the true probability lies between P_0 and P_2

They are simply told that the probability of 'making an error' <10%. In effect, these materials have adopted a nomenclature that says that for the interval of true proportions over which the probability of type II error is $>\beta$, they will not call a type II error an error. This confusing logic is not described to the trainees, so they are left to draw their own straightforward conclusions. Unfortunately, they probably find their conclusions to be reassuring.

Concern 2c: The trainees probably come away with the sense that small sample studies can be very powerful

If we adopt the intuitive definitions of errors, and focus on a single threshold, then we might infer that the study design can reliably discern between situations where $P_d = 79\%$ and $P_d = 81\%$ with only 10% error rates. Such a design is depicted in Figure 6. We feel that it is likely that the trainees come away with the feeling that a study where $n = 19$ and $d^* = 13$ has the type of power that can only be achieved with $n = 2800$ and $d^* = 2240$.

Concern 3: The training materials use language that obscures the bias of the null hypothesis

If a study finds more than d^* persons with the trait of interest, then the instructions that accompany Table 1 in Valadez and Devkota[4] say that the supervisor should judge the district as having 'reached the threshold'. In the language of hypothesis

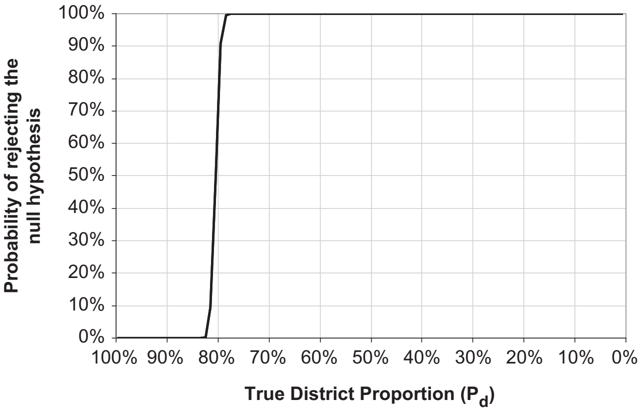

Figure 6 Operating characteristic curve for a $n = 2800$, $d^* = 2240$ design

testing, the supervisor is encouraged to 'accept' the null hypothesis. This is a fundamental error in hypothesis testing and we find this language to be very misleading. The active tone of the phrase, 'reached the threshold' makes it sound like this is a conclusion that requires strong evidence, when, in fact, it is the default conclusion—the null hypothesis.

Furthermore, the notion of stating that the population has 'reached the threshold' when the sample proportion falls below the threshold, can lead to some nonsensical conclusions when the data from multiple districts are aggregated. Imagine a region with several dozen districts, each of which produces samples where exactly 13 out of 19 persons are vaccinated. The Health Ministry press release might read something like this:

> 'Every district has reached the vaccination threshold of 80% and the region's overall vaccination rate is 68.4% ± 3%.'

We are not advocating the use of complicated statistical verbiage to phrase LQAS conclusions. We realize that precise descriptions of results that fail to reject the null hypothesis often contain double-negatives and tend to be difficult for laypersons to parse. Instead, we are saying that study designs should make clear what conclusion they will draw by default (the null hypothesis) and what conclusion requires strong evidence.

Concern 4: The use of the word 'optimal' reinforces these misunderstandings

Table 1 in Valadez and Devkota[4] uses the word 'optimal' in its title. The training materials available on the Internet state that the 'optimal size for cluster sampling projects = 19'. Appendix 3 of the trainer's manual circles some study designs and describes them as 'optimal'.[12]

Persons who are quantitatively adept, know that in order to be meaningful, the word 'optimal' should be accompanied by a list of objectives and constraints. We fear that trainees who are not quantitatively adept are likely to hear the word 'optimal' as 'optimal for me'.

Recommendations

In light of these concerns, we make the following recommendations for persons designing and reporting LQAS studies, for editors reviewing papers or reports that report LQAS work, and for persons who develop LQAS training materials.

(1) We strongly recommend that the null hypothesis should always protect the population at risk.

(2) Regardless of the direction of the null hypothesis, LQAS study designs should always be described in a way that clearly states which conclusion requires only weak evidence and which one requires strong evidence. LQAS training materials should make it clear that each district's LQAS hypothesis test either protects the people, or is biased against them from the start. This is a fundamental property of any hypothesis test that controls the probability of type I error first, and then minimizes the probability of type II error. This inherent feature should be emphasized to LQAS designers and trainees, and we believe this can be accomplished with a simple quotient, and without using statistical jargon.

(3) Specifically, compute the quotient $(d^*)/n$ and compare it to P_0. If you wish to confidently conclude that $P_d > P_0$, then your test should require a sample proportion that is $>P_0$. Otherwise, the test lacks face validity. To confidently conclude that $P_d > 80\%$, a test should require a sample proportion that is >80%.

(4) If 'error rates' are listed, then the term 'error' should be defined clearly.

(5) LQAS training materials should develop the concept of 'error' with the trainees. We suggest that the simple term 'error' be reserved for the simple definition that trainees are likely to infer naturally. If we conclude that a district has reached the threshold when it has not, we have made an error. If we conclude that the district has not reached the threshold when it has, we have made an error. Small sample studies will have high probabilities of making type II errors when $P_0 < P_d < P_2$.

(6) Trainees and LQAS designers should be made to understand that small sample studies have low power. They will frequently result in classification errors. They can only be 'optimal' in a big picture, bureaucratic sense of trading off time and resources and they are blunt

instruments at best for classifying whether or not individual districts have reached a particular threshold.

(7) Study designers should understand that if they need more power for decision-making at the district level, then they will need to adopt a design with a larger value of n, either in the form of a single-sample design or a double-sample LQAS design that collects more information if the first sample is inconclusive.[3]

(8) The word ‘optimal’ should not be used without being clearly defined.

Conclusion

LQAS studies can accomplish important goals at relatively low cost, but as the title of our article states clearly, we urge users of LQAS study designs to beware. In order to be credible, sampling designs must be statistically sound and authors who describe LQAS work should make their assumptions and implications perfectly clear. We are especially concerned that life-giving resources may be prematurely withdrawn from needy populations based on faulty conclusions. We feel strongly that study designers have a responsibility to protect the population at risk. For the sake of those vulnerable populations, we recommend that the existing training materials be thoroughly overhauled, and that authors of LQAS manuscripts and reports be held to a new high standard of clarity and face validity.

Conflict of Interest: None declared.

References

[1] Robertson SE, Valadez J. Global Review of health care surveys using lot quality assurance sampling (LQAS) 1984–2004. *Soc Sci Med* 2006;**63:**1648–60.

[2] Lemeshow S, Stroh G. Quality assurance sampling for evaluating health parameters in developing countries. *Surv Methodol* 1989;**15:**71–81.

[3] Lemeshow S, Taber S. Lot quality assurance sampling: single- and double sampling plans. *World Health Stat Q* 1991;**44:**115–32.

[4] Valadez JJ, Devkota BR. Decentralized supervision of community health programs: using LQAS in two districts of southern Nepal. In: Rhode J, Wyon J (eds). *Community Based Health Care: Lessons from Bangladesh to Boston*. Boston: Management Sciences for Health, 2002. pp. 160–200.

[5] Valadez J, Weiss W, Leburg C, Davis R. *Assessing community health programs: A participant’s manual and workbook: using LQAS for baseline surveys and regular monitoring*. London: Teaching Aids at Low Cost (TALC), 2003.

[6] CORE Monitoring and Evaluation Workgroup LQAS Online Series. 2006. Available at: http://www.coregroup.org/conf_reg/lqas_series.cfm, (Accessed 2 January 2009).

[7] LQAS Lecture #1. 2006. Available at: http://www.coregroup.org/conf_reg/LQAS_Lecture_1.pdf. (Accessed 2 January 2009).

[8] Valadez JJ, Weiss W, Leburg C, Davis R. *Assessing Community Health Programs: A Participant’s Manual and Workbook: Using LQAS for Baseline Surveys and Regular Monitoring*. *Monograph on the Internet*. 2002. Available at: http://www.coregroup.org/working_groups/LQAS_Participant_Manual_L.pdf. (Accessed 2 January 2009).

[9] Dodge HF, Romig HG. *Sampling Inspection Tables*. New York: John Wiley & Sons, 1944.

[10] Dodge HF, Romig HG. *Sampling Inspection Tables*. 2nd edn. New York: John Wiley & Sons, 1959.

[11] Fitch DJ. Is the Valadez evaluation method based on Dodge and Romig? *Proceedings of the International Statistical Institute*. Portugal, Lisboa: International Statistical Institute, 2007.

[12] Valadez J, Weiss W, Leburg C, Davis R. *Assessing community health programs: A Trainer’s Guide: Using LQAS for Baseline Surveys and Regular Monitoring*. *Monograph on Internet*. London: Teaching Aids at Low Cost (TALC); 2003. Available at: http://www.coregroup.org/working_groups/lqas_train.html (Accessed 2 January 2009).