Cry Toxins in Transgenic Plants Have Direct Effects on Natural Enemies in the Laboratory

To the Editor:

Previously, we demonstrated via meta-analysis that Cry toxins and proteinase inhibitors that are used in GM plants have nonzero direct and indirect effects on arthropod predators and parasitoids in the laboratory (Andow et al. 2009, Lövei et al. 2009). Our purpose was not to analyze the risks associated with any particular GM plant, but to measure whether there are detectable nonzero effects and the implications for assessing risk. These conclusions were criticized by Shelton et al. (2009b), and clarify the theory underlying our statistical tests.

Statistical Issues: Theory and Methodological Details. There remain some misunderstandings about the structure of our statistical model for meta-analysis and the nature of the null hypothesis. Meta-analysis can be conducted in two fundamentally different cases. In Case I, which is most common in medical meta-analysis, data are aggregated from several studies measuring the same response on the same species by using the same laboratory protocols. Meta-analysis theory assumes that the real response underlying the experimental observations is either positive, negative, or zero, but by necessity, it has a single value. Under Case I, the null hypothesis is that if there is one real response, \( \mu \), then the normalized data have a standard normal distribution with a mean equal to the single real response value,

\[ H_0: g = N(\mu, 1) \text{, for some } \mu \]

where \( g \) is the normalized data. If this null hypothesis is rejected, then there is not one real response value, and the evidence originates from >1 populations with different real response values (i.e., the original population is not homogeneous). However, this is irrelevant to the present case.

In Case II, which applies here and is common in ecological meta-analysis, the aggregated data are similar responses measured on different species by using similar, but not identical experimental protocols. In this case, there are multiple underlying real responses, because each species can have different means, and the means for each protocol can be different (Andow et al. 2009). Indeed, it would border on the absurd to assert that all species have the same underlying response. The derivation of the null hypothesis under Case II is different from that for Case I. First, the null hypothesis assumes that there are multiple real response values. Next, we know from the theory for Case I that the distribution of the normalized data have a standard normal distribution around their real response value, i.e., \( g_{rse} \sim N(\mu_{rse}, 1) \) for each response \( r \), species \( s \), and experiment condition \( e \), where the \( \mu_{rse} \) may differ for each measured response, species, and/or experimental condition. The Case II null hypothesis is developed by assuming that all of the different real response values are zero,

\[ H_0: \mu_{rse} = 0 \text{ for all } r, s, \text{ and } e, \text{ then } g_{rse} \sim N(0, 1) \text{ for all } r, s, \text{ and } e. \]

The null hypothesis states that the evidence will be distributed as a standard normal around a mean of zero (not an arbitrary \( \mu \) as in Case I). If \( g_{rse} \) is not normally distributed around zero, then the null hypothesis is false, and we must conclude that some of the real responses (means) are nonzero. For Case II, the analyst cannot conclude that there are multiple populations (as suggested by Shelton et al. 2009b), because this assumption is already part of the null hypothesis.

At least three statistical measures (Hedges and Olkin 1985), Cochran’s Q, (called “heterogeneity analysis” by Shelton et al. [2009b]), Fisher sum of the log \( P \) values, and a log-linear \( \chi^2 \) test for goodness-of-fit (GOF) may be used to determine whether evidence is normally distributed. Cochran’s Q and Fisher sum of the log \( P \) values are statistically much more powerful than a GOF test (Hedges and Olkin 1985). Shelton et al. (2009b) suggested that we should have used Cochran’s Q, but we elected to use the less powerful GOF test, because it is robust to extreme values and statistical error in the data and is less likely to reject the null hypothesis than Cochran’s Q. Thus, this criticism does not invalidate our conclusions; it strengthens them. We plan to complete an analysis of our data using the more sensitive Cochran’s Q and make the data publicly available.

Previously we stated that most of the studies in our sample had sufficient replication (\( \geq 20 \) observations) that our use of Hedges’ \( g \) to normalize the evidence was sufficient. We note that the number of observations per study is the number of independent individuals or groups of individuals in the study. Because pseudo-replication is not uncommon in entomological publications (Hurlbert and Meikle 2003), we closely examined the methodological descriptions in each of the papers included in our analysis. In particular, we noted if individuals had been tested individually or in groups. In the first case, individuals would be independent units of replication, and in the second case, the number of groups of individuals would be the appropriate unit of replication. Based on the methods
in the papers we reviewed, 96% of the evidence had \( \geq 20 \) independent observations, which is what we expect from well-conducted laboratory studies.

We would also like to rectify a notational error in the earlier description of our statistical methods in Andow et al. (2009). Using the same notation as before (Andow et al. 2009), we calculated

\[
\frac{\bar{x}_t - \bar{x}_c}{\sqrt{\frac{n_t n_c}{n_t + n_c}}}
\]

where \( SD_p \) is the pooled standard deviation, and \( \bar{x} \) is the mean and \( n \) is the sample size of either the treatment \( t \) or the control \( c \). We divided [1] by [2], which is equal to scaled \( g \) and is distributed generally as a noncentral \( t \)-distribution. We erroneously called [2] a standard error in Andow et al. (2009), but our calculations used [2] were correct. The data in our analysis are counts of the values of scaled \( g \) broken into five intervals, \( g < -2, -2 \leq g < -1, g = 1, 1 \leq g < 2, \) and \( 2 < g \).

**Revisiting Statistical Independence.** Shelton et al. (2009b) suggested that there would be positive correlations among responses measured on different instars of the same individuals (e.g., development rate measured on subsequent instars) so that the observed statistically significant non-normal distribution was an artifact of the positive correlation. We noted previously that the actual data may not be correlated—indeed, all of the original studies assumed that their data were independent, and we showed that the data are not strongly correlated (Lövei et al. 2009). Clearly, if there are insignificant positive correlations, Shelton et al.’s concern is invalid. If factors other than a positive correlation cause the nonnormal distribution, this would also invalidate the concern. We suggest that future investigators should examine potential intra-individual correlations in all their data.

Their larger point, however, is that there are no nonzero effects of Cry toxin on natural enemies. We show that even if there were positive correlations among the same responses measured on different instars, and these correlations, by themselves, generated the nonnormal distribution of responses, the existence of nonzero direct effects of Cry toxins on natural enemies remains valid. If the instar-specific responses are positively correlated, then a logical consequence is that the total response, measured over all of the instars, must also be nonnormal. Therefore Cry toxins have nonzero effects on the total response, which refutes their larger point.

This result can be explained as follows: We know that the aggregated instar-specific responses are non-normal. This implies that at least one of the instar responses by itself is significantly nonnormal, or several of them are almost nonnormal. In either case, because of the positive correlation among instar-specific responses, there will be several instar responses that are almost nonnormal and they will be skewed in the same direction. Therefore, when added together into the total response, they will generate a distribution that is skewed even more than any of the individual instar responses. Consequently, the total response will be nonnormal, and from this it follows that there are nonzero responses to Cry toxins.

**Significance.** We believe that one of our main findings, that Cry toxins have nonzero direct effects on natural enemies in laboratory studies stands up to the criticism by Shelton et al. (2009a,b). This finding suggests that classic Tier 1 tests for ERA may not be appropriate for genetically engineered plants. These are acute toxicity tests conducted in the laboratory on a surrogate species (Romeis et al. 2008). The Tier 1 test is promoted as a worst-case test that is designed to guard specifically against false negatives, so that if no effect is found, then the risk can be concluded to be zero. Romeis et al. (2008) suggest that in this case, an ERA could be stopped without any additional measurement in the laboratory or field.

Our finding indicates that some Tier 1 laboratory tests of Cry toxins on natural enemies had reported no effect of a Cry toxin when actually there was an effect. Although it is widely known that anytime a researcher fails to reject the null hypothesis, there could be a type II error (erroneous failure to reject), our point is different. We conclude not that there could have been type II error, but that there has been type II error. A false belief of no effect, when an effect does exist, implies that some ecological risks will not be properly measured because of a premature conclusion of no risk. As we have suggested elsewhere (Andow and Hilbeck 2004, Lövei et al. 2007), another approach to ERA for transgenic crops (and perhaps for pesticides and other environmental toxins) should be considered, and we hope that our work helps stimulate such developments. We believe that this is in the interest of all stakeholders concerned.

**Acknowledgments**

We thank Drs. M. Nakai and E. Yano for discussions that helped clarify our presentation, Drs. A. Rendahl and S. Weisberg for their helpful statistical advice, and an anonymous reviewer for spotting an error in the statistical description. We thank the Bellagio Center of the Rockefeller Foundation for providing support to finish this letter.

**References Cited**


Errors in Logic and Statistics Plague a Meta-Analysis (response to Andow and Lövei 2012)

TO THE EDITOR:

As we noted previously (Shelton et al. 2009a,b), we strongly believe in the power of meta-analyses to help advance our collective understanding of the potential risks of Bt crops for nontarget organisms by identifying negative, neutral, and positive effects of the technology in both laboratory and field studies. Although we agree on this point, it is equally important that such studies do not contain errors in logic or statistics. We acknowledge that Andow and Lövei (2012) have corrected a statistical error in their previous publications (Lövei et al. 2009), but point out another statistical error (see below) in their latest letter. However, more important than these statistical errors, we question their continued attribution of hazard to a protein rather than, more accurately, to poor prey or host quality. We believe this is an error in logic. Therefore, we strongly oppose the latest statement by Andow and Lövei (2012) that their conclusions of detectable non-zero effects of Cry proteins on nontarget organisms were “...criticized by Shelton et al. (2009a,b) on statistical grounds.” They missed the point again. Our primary criticism then and now is that they continue to ignore prey and host-quality effects and the ecological context in their effort to inform risk assessment.

Besides the fault in logic of ignoring prey or host quality, we also take issue with the statement in their latest letter (Andow and Lövei 2012): “We conclude not that there could have been type II error, but that there has been type II error.” Given the nature of statistical hypothesis testing, such certainty is simply not possible. The only thing they can conclude is that they rejected the null hypothesis. This could just as likely represent a type I error on their part as it could a type II error by every other meta-analysis conducted to date—an unlikely conclusion implied by their statement. Whether their statement is an error in statistics, logic, or both is a matter of debate, but it is an error. We also would like to point out that unlike other published meta-analyses that have made the underlying database accessible to the scientific community (e.g., Wolfenbarger et al. 2008, Duan et al. 2010), the database supporting the report by Lövei et al. (2009) was not made accessible and therefore cannot be measured for statistical veracity or other important criteria by interested parties. However, in the end, the statistical issues debated exhaustively by both parties are irrelevant if there are errors in the logic of attributing hazard to a protein rather poor prey or host quality, as they did in their study (Lövei et al. 2009). We strongly believe that the analysis and conclusions stated by Lövei et al. (2009) do not provide evidence for toxic effects of Bt Cry proteins on natural enemies.

We agree that tritrophic laboratory studies have reported adverse effects of Bt-transgenic plants on natural enemies. However, these effects must be regarded as prey-quality effects rather than toxic effects of the plant-expressed Cry proteins, because these tritrophic studies have used Bt susceptible insects as hosts and prey for the natural enemy. When a host is susceptible and ingests a Bt protein, its quality is reduced and when it is fed to a natural enemy this might result in a negative effect on the natural enemy. But it is not the protein itself that has the effect! A careful reading of the literature shows that when effects have been observed, it was the poor quality of the Bt-susceptible host that was responsible for the observed findings (Romeis et al. 2006, Naranjo 2009). This has been verified in tritrophic studies conducted with Bt-resistant or nonsusceptible herbivores. The literature has shown that allowing Bt-resistant hosts to ingest Bt proteins and then feeding the hosts to natural enemies (both predators and parasitoids) has revealed no effects on the natural enemies (Table 1). Likewise, the literature has shown that exposing natural enemies to nonsusceptible prey that have fed on Bt proteins has revealed a lack of effect (Table 1). Meta-analyses (Naranjo 2009) have further demonstrated that, with removal of prey and host quality as a confounding factor, the effects of Bt proteins are either neutral or


D. A. Andow
Department of Entomology and Center for Community Genetics
University of Minnesota
St. Paul, MN

G. L. Lövei
Department of Agroecology
Aarhus University
Flakkebjerg Research Centre
Slagelse, Denmark