A/B Testing for Industrial companies

Anjali Agrawal
Computer Science and Engineering
B.Tech.
of SRM Institute of Science and Technology
Chennai, India
aa9552@srmist.edu.in

Dr. S. Prabakeran
Department Of Networking And Communications
of SRM Institute of Science and Technology
Chennai, India
prabakes@srmist.edu.in

Preeti Vijayan
Technology Consulting - BI & Analytics
Tiger Analytics
Chennai, India
preeti.vijayan@tigeranalytics.com

Abstract—

User-intensive software, like mobile applications and websites, heavily depends on interactions with many users and an unknown population. With the internet connectivity on such software, the website allows evaluating ideas and innovations using continuous experiments like A/B tests, split tests. We aim to study the application of A/B testing in various industrial contexts. We will also present a brief study on different statistical tests for different assumptions and solve them using the Bayesian algorithm. Randomized algorithms are used to address various software engineering problems. This type of algorithm gives different results with every run for the same problem instance. Therefore, a statistical test is important to prove the conclusions derived from the data.

Keywords— A/B testing, split testing, statistical tests, Bayesian Testing

I. INTRODUCTION

A/B test (controlled or randomized experiment) is the standard method to make data-driven decisions. It is the process of testing two or more versions of marketing assets, to identify which one performs better. It takes the output from digital marketing through the collection of data. By analyzing the result of each version, we get an understanding of what works and what doesn't. It increases the chance of customer conversion and user engagements with the product.

1. **A/B test vs Split test:** The main difference is A/B testing involves comparing two different versions of marketing assets based on changing one element, such as the CTA text or image or color on a landing page. While split test involves comparison of two distinct designs.

   Champions, Challengers, and Variations in A/B test: Champion is a marketing asset like a webpage that we think will perform well or have performed well in past. Challenger is a variation of the champion with one different element. After the A/B test is done, either we get a new champion, or we find that the first variation performed better.

Structure of Controlled Experiment and A/B Testing

Elements of an experimentation system: The basic setup is to evaluate an asset with two levels: control, and test. A control version is a default case and a test version is one where variation is been experimented with. The test is extended and is commonly referred to as A/B/n split tests. A multifactor experiment is referred to as multivariable or multivariate.

The high-level framework of an A/B experiment is depicted in Figure(1) below. In a practical scenario, any proportion can be assigned to the test and control, but 50 percent gives the test the most statistical power, and it is recommended to power the trials at lesser percentages after a ramp-up phase to check for severe mistakes.

Generally, the analysis examines whether the statical distribution is different for test and control. In practice, we check the two means are equal or not. For the same, the effect of the test is defined as

$$E(B)= X'B - X'A$$

Where X is a metric of interest and X'B is the mean of the test. The percentage change is reported with a suitable interval.

The two essential elements of control are extraneous factors and randomization. Any factor that affects the examination is either an experiment factor or non-experiment factor. The non-experiment factor might be kept constant, blocked, or randomized. Holding a factor constant can influence external validity and is hence not advised.

For example, if weekend days are known to be different from weekdays, you might experiment exclusively on weekdays, although it would be preferable to include whole weeks in the experiment for higher external validity. Blocking can minimize variation compared to randomness and is recommended when the experimentation units

within each block are more homogeneous than the units between blocks.

If the randomization unit is a user page visit, for example, blocking by weekend/weekday might minimize the variation of the impact magnitude, resulting in improved sensitivity. Because many external factors alter with time, it is crucial to randomize throughout time by running the control and test(s) concurrently with a set percentage to each throughout the experiment.

Controlling a non-treatment component ensures that it has an equal influence on both the control and the test, and so has no effect on the estimate of the testing effect.

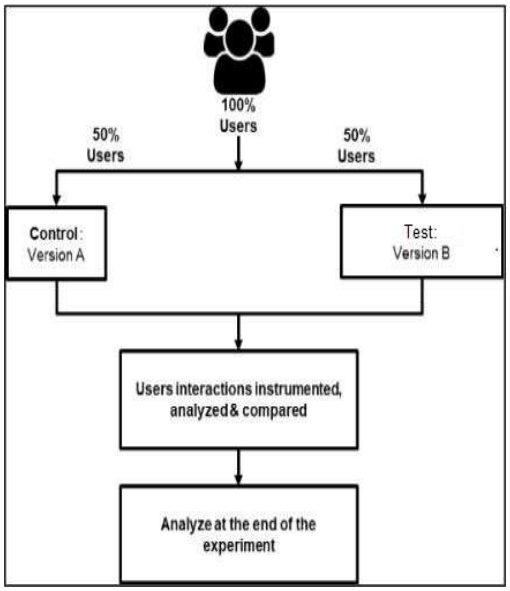


Figure (1): The high-level framework of an A/B experiment

II. EXPERIMENTS

1. Test Requirements

The main objective of any firm is to maximize revenue. Here I will be demonstrating my personal work experience for a security company. They want to modify the official website of the company and add a few features to increase the revenue. Figure (2) illustrates the test. Currently, the company gives a complete page with normal blue and white shades when the user opens the app. They want to change the page structure and give an overlay of red color while auto launching the app in users' systems.

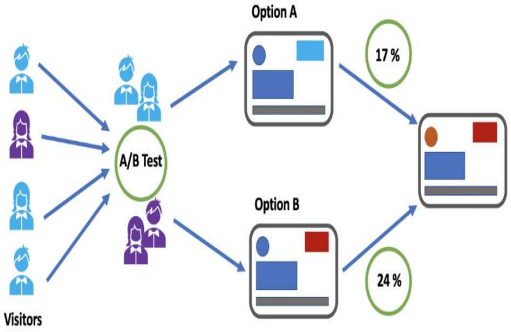
They want to check if this transition affects renewal count, or customers leaving the product. The measure of interest is lifted in total conversion, lift in apps uninstalled, customer retention, and weekly conversions.

In the execution of this experiment, users are assigned to option A or option B randomly. The only difference between the two is one is getting overlay automatically if they switch on the pc and the second is they manually check the app. So, conversion is based on the condition that either revenue increased by auto-launch or uninstallation increased because of it. To determine the relevance of the test, we use the Bayesian model of A/B testing. We determine the relevance of difference; Frequentist inferences are used.

Some common questions answered by the Bayesian approach are:

The probability that test is better than control, total uplift in conversion if we roll-out test, and risk of switching to test from control.

In this paper, we define the approach for an analyst but without sharing actual data, as the code of conduct of the company. Here, few clients are the population. Clients are selected based on certain criteria; the total lift should be greater than the old one is the success of the test. Clients engaging on the app or clicking on the overlay is the area of interest.



Figure(2) Illustration of basic experimentation with two variants

2. Understanding Concept

We follow "Bayesian Data Analysis" by Gelman for specific details like formulas and concepts. Bayesian statistics is based on the principle of probability statements. It also states how to update probabilities after obtaining new data. In Bayesian statistics, we have a concept called conjugacy, which means posterior distribution P(A|X) is in the same probability distribution family as prior distribution P(A). The prior is then called a conjugate prior.

$$P(A|X) = \frac{P(A)P(X|A)}{P(X).}$$

It describes the dependency of the posterior distribution P(A|X) of a parameter A after seeing data X with prior distribution P(A) of parameter A. To describe this model we need six distributions, which are: The binomial distribution, The Beta distribution, The multinomial distribution, The Dirichlet distribution, The exponential distribution, and The Gamma distribution.

We have different experiment scenarios like comparing variations where the end-user has one option to choose, comparing variations where the end-user has many options to choose, comparing variations with many options but we observe final lift only. In our case, we will work on the third model that is aggregated model.

3. Aggregate Model

In this process, aggregated conversion, and the sum of revenue generation per variant are observed. This happens because of too many options within a variant or because of the data generation process. Some of our tests have more than 20 different options to choose from, for all required conditions we formalize the scenario as follows.

For a given variant a, there are K_a unknown options. N_a is total visitors and C_a is total conversions. The individual revenue per instance V_a is unknown. Executing the experiment results in collected data C_a, and the aggregated revenue S_a overall C_a successes and implicitly over all unknown K_a options, i.e., $S_a = \sum C^i_{j=1} S_{ij}$ with S_{ij} the revenue of each success j. To get an estimate of the revenue per visitor V_i we model the average revenue per visitor V_i' given the observed aggregated revenue S_i. For that, we assume that S_{ij} follows an exponential distribution

$$P(S_{ij}|V_{i'}) = Exp(V_i')$$

where V_i' is the scale parameter of the distribution. The exponential distribution means that lower revenue has more probability of occurrence than higher revenue. This assumption fits our observation of the visitors' money spent curve on our website. The prior distribution of V_i' is modeled as

$$P(V_i) = Gamma\ (\alpha_i\ ,\ \beta_i)$$

Taking advantage of the conjugacy relationship between exponential and gamma distributions, we compute the posterior distribution as

$$P(V_i|S_i) = Gamma\ (\alpha + C_i\ ,\ \beta_i + S_i)$$

The expected value per client per variation is $A_i * V_i$. Since we have the posterior distributions for both, A_i and V_i, we take n random samples from $P(A_i|X_i)$ for the conversion rate Y_i, n random samples from $P(V_i|S_i)$ for the revenue.

4. Making decisions for the test

All the tests have a feature defined, which decides if the test was successful or not, or which variation will give better output in terms of conversion rate, uninstallation rate, click-through rate, etc.

a. Probability of getting the best result: After running a test with many variations, we want to derive the calculation for the best variation and which one to be implemented. Given a set of posterior samples Y_i of the measure of interest from *i* different variants, the probability for the best is defined as the probability that a variation has a higher measure in comparison with all other variations. The probability that Y_1 is better than Y_2 is the mean of:

$$P\ (Y_1 > Y_2) = [y_{1j} > y_{2j}\ |\ i \text{ and } j \in \{1, \ldots, n\}]$$

To find the best variation, we analyze all combinations and select the one with the highest probability.

b. Expected lift

After getting the best variant, we also need to check the increase in the measure of the matrix we expect. Given a set of posterior samples Y_i, the expected uplift of choosing Y_1 over Y_2 is defined as the mean of the percentage increase:

$$U\ (Y_1, Y_2) = [\ (y_{1i} - y_{2i})/\ y_{2i}\ |\ i \text{ and } j \in \{1, \ldots, n\}]$$

We want uplift compared to control variation only, we calculate it for all test variations against each other.

c. Expected loss

Suppose variation 1 has the highest probability to be the best but is smaller than 1. Then, there is still a chance that the other variation is the true best performing one. In such a case we want to know the risk of opting in Variation 1. Given a set of posterior samples Y_i, the expected loss when choosing Y_1 over Y_2 is the mean of:

$$L\ (Y_1, Y_2) = [\max\ ((y_{2j} - y_{1j})/y_{1j}\ ,\ 0)\ |\ i \text{ and } j \in \{1, .., n\}]$$

Like the expected uplift, the expected loss is calculated between the test variant and the control.

5. Common mistakes done in A/B test

In this fast-moving world, companies look for a first-user advantage. The most common mistake evolves out of impatience. Many managers don't let the tests run for the required duration as they want to plan quickly. This is solved by a different type of A/B testing called "real-time optimization". In this, we use different algorithms to adjust results with time duration. Because of this adjustment, we may get a different result, if it runs for the full course. The second mistake is having too many metrics. If the analyst is looking at hundreds of metrics at the same time, they may risk getting into spurious correlations. More the number of tests, chances of getting accurate results to increase, but it also increases the risk of confusion as we may get many different outputs. We have more complex tests that are better than the A/B test but in A/B testing everything happens so quickly that if something doesn't work, we can go to the old tactic.

6. Architecture Alternatives for Experimentation

Web-based controlled experiments: an overview and a practical guide (Kohavi et al. 2009) presents an overview of several architectural options. The randomization algorithm, the assignment method (i.e., how the randomly allocated experimental units are given the variations), and the data path are the three essential components of an experimentation capability (which captures raw observation data and processes it). Tang et al. (2010) provide a thorough overview of Google's experimental infrastructure.

We recommend running A/A tests frequently to ensure that the experimental setup and randomization process is operating effectively. An A/A test, also known as a null test (Peterson 2004), challenges the experimental system by randomly allocating users to one of two groups but exposing them to the identical experience. An A/A test may be used to gather data and analyze its variability for power calculations, and (ii) test the experimental system (the null hypothesis should be rejected roughly 5% of the time at a 95% confidence level)

7. Experiments are being planned.

Several components of experiment preparation are critical, including calculating acceptable sample size, acquiring the proper metrics, following the right people, and selecting a randomization unit.

The sample size. The sample size is defined by the percentage of users accepted to the experiment variations (control and treatments) as well as the length of the trial. As an experiment continues, more visitors are accepted to the versions, resulting in larger sample sizes. Experimenters can determine the percentage of visitors who are in the control and treatment groups, which impacts how long the experiment must run.

Several writers (Deng et al. 2013; Kohavi et al. 2009) have addressed the question of sample size and experiment time to attain acceptable statistical power for an experiment, where statistical power is the chance of detecting a certain effect when it occurs (technically, the probability of correctly rejecting the null hypothesis when it is false). In addition to arranging an experiment with enough power, it is recommended to experiment for at least one week (to catch a full weekly cycle) and then for numerous weeks beyond that. When "novelty" or "primacy" effects are suspected (i.e., the treatment's early effect is not the same as the long-term effect), the experiment should be performed long enough to estimate the treatment's asymptotic effect. Finally, analyzing the influence on a high-variance statistic, such as loyalty (sessions/user), would often need a greater number of users than other metrics (Kohavi et al. 2012).

8. Metrics, Observations, and the OEC

Gathering observations (i.e., recording events) to compute the appropriate metrics is crucial to effective experimentation. When practicable and economically practical, acquire as many observations as possible that contribute to addressing prospective issues of interest, whether O user or performance related (e.g., latency, utilization, crashes). We propose computing a large number of metrics from the data (e.g., hundreds) since they might yield unexpected insights, albeit care must be made to accurately assess and adjust for the false-positive rate. While having a large number of indicators is beneficial for gaining insights, judgments should be made using the Overall Evaluation Criterion (OEC). The OEC is described in Tenet 1 previously.

9. Trigger:

Some therapies may apply to all website visitors. However, for many trials, the introduced difference is only important for a

subset of visitors (e.g., a change to the checkout process, which only 10 percent of visitors start). In these circumstances, only visitors who would have seen a difference in one of the versions should be included (this commonly requires counterfactual triggering for the control). Some designs overtly or through lazy (or late-bound) assignments force users to participate in an experiment. In any situation, the idea is to examine only the demographic subset that was possibly impacted. Triggering minimizes the variability in treatment effect estimates, resulting in more exact estimates. Because the diluted impact is frequently of interest, it can be diluted.

10. **Randomization Unit.**

For a variety of reasons, most trials employ the visitor as the randomization unit. First and foremost, for many of the modifications being evaluated, it is critical to provide the consumer with a consistent online experience. Second, most experimenters assess measures like sessions per user and clicks per user at the user level. In an ideal world, the experimenter would use a genuine user to randomize, however, in many unauthenticated sites, a cookie stored by the user's browser is utilized, hence the randomization unit is the cookie. In this situation, the same user will appear to be a separate user if she visits the site with a different browser, device, or after deleting her cookie during the trial. The next part will go through how the randomization unit chosen influences how different metrics should be analyzed. The randomization unit can also influence the test's power for specific measures. Deng et al. (2011), for example, demonstrated that randomization at the page level may considerably reduce the variation of page-level metrics, although user metrics cannot be obtained in such instances. Spillover effects in social networks contradict the traditional non-interference assumption, necessitating novel techniques such as clustering.

11. **Experiment Analysis**

If an experiment is properly carried out, the analysis should be a simple application of well-known statistical procedures. Of course, this is preferable to attempting to recover from a flawed experimental design or execution.

12. **Decision-making:**

The typical hypothesis-testing process, assuming the normal distribution if the sample size is sufficient, is a popular approach to determining if the treatment is better than the control. When normalcy cannot be assumed, data transformations and nonparametric or resampling/permutation procedures are used to establish how odd the observed sample is under the null hypothesis (Good 2005). A p-value of the statistical test is frequently provided as evidence when conducting a test to determine if the treatment had an impact or not (e.g., a test to determine whether the treatment and control means are equal). More specifically, the p-value is the chance of obtaining an effect that is equal to or more severe than the one seen, assuming that the null hypothesis of no effect is correct. Another option is to apply Bayes' theorem to compute the posterior probability that the therapy had a positive impact vs the odds that it had no effect (Stone 2013).

13. **Units of analysis:**

Metrics can be defined using various analytic units, such as user, session, or other suitable bases. An e-commerce site, for example, would be interested in metrics such as revenue per user, revenue per session, or revenue per purchaser. Simple statistical procedures (such as the t-test and variations) apply to any measure that has a user as its analysis unit if users are the unit of randomization since users may be considered independent. However, if the analysis unit and the randomization unit are not the same, the analysis units may not be deemed independent, and alternative methods must be employed to determine

standard deviation or compare the treatment to control. When the analysis unit and the randomization unit are not the same, two often-used approaches are bootstrapping and the delta method

14. **Reduced Variance:**

One method for increasing power is to increase the sample size. Online researchers, on the other hand, are always seeking ways to boost the power of their studies while decreasing, or at the very least not lengthening, the length of the tests. Covariates such as pre-experiment user metrics, user demographics, location, equipment, software, connection speed, and so on can help with this. (Deng et al. 2013) provided an example of how utilizing solely the pre-experiment metric values for the users might result in a 50% decrease in variation for a measure.

15. **Diagnostics:**

Every experimentation system should have certain diagnostic tools to ensure the reliability of the experimental data. Graphs illustrating the number of users in each version, metric means, and treatment effects over time will assist the researcher in identifying any unforeseen difficulties or disruptions to the experiment. Furthermore, diagnostic tests that generate an alert when an expected condition is not satisfied should be included. The "sample ratio mismatch," or SRM, is an important diagnostic test. A basic statistical test is used to determine whether the actual percentages for each variant are near enough to the anticipated percentages. We've discovered that for online trials, this one diagnostic is typically the "canary in the coal mine." There are several methods for an experiment to skew the number of visits to one version or another, and many of them will result in a significant bias in the treatment effect. Another typical beneficial test is to ensure that the performance, or latency, of the two versions, is comparable as expected. In certain circumstances, the treatment may be slower owing to caching concerns (e.g., cold start), or if the variants are imbalanced (e.g., 90/10 percent), a shared resource such as an LRU cache (Least Recently Used) will give the bigger variation an advantage. When an experimentation platform supports overlapping experiments, a diagnostic to detect interactions between overlapping experiments is also useful. When an alarm or graph shows a possible problem, the researcher should look into it to find out what's causing it.

III. ACKNOWLEDGMENT

I am grateful to the members of the organizing committee of the Third International Conference on IoT Based Control Networks and Intelligent Systems (ICICNIS 2022), in particular Professor Prabakeran S, for his guidance on the interaction between statistics and software testing and presenting my work, both through a presentation at the conference and this article. The contents of this contribution are fully from the author's perspective but are based on several ongoing research collaborations and industry working models, as reflected by the bibliography of this article. I sincerely thank all collaborators for their substantial inputs in jointly authored research publications, and beyond that to everyone who have played a major part in shaping my research activities and views on interesting research challenges

IV. CONCLUSION

Data-driven decisions have the greatest impact on output and productivity, which is why they are so important in running a firm. Companies perform experiments regularly to aid decision-making. In this paper, we describe how systematic evaluation is done, and how the results are analyzed in software engineering. The online connectivity of clients' software has increased the scope of online experimentation. This makes experimentation much more scalable and efficient and allows quick evaluation. For easy interpretation, we use the Bayesian approach. We also described the three most common scenarios of evaluation and solved one for it. We also conclude that continuous experiments and the A/B test can be an active topic to research as there are many research gaps to fill.

V. REFERENCES

The below provided are utilized references to get information:

[1] G. Eason, B. Noble, and I. N. Sneddon, “On certain integrals of Lipschitz-Hankel type involving products of Bessel functions,” Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955.

[2] A. Denise, M.-C. Gaudel and S.-D. Gouraud email= {denise,mcg,gouraud}@lri.fr L.R.I., Université Paris-Sud, 91405 Orsay Cedex, France. A Generic Method for Statistical Testing. Proceedings of the 15th International Symposium on Software Reliability Engineering (ISSRE’04) 1071-9458/04 IEEE.https://www.lri.fr/~bibli/Rapports-internes/2003/RR1378.pdf.

[3] Nikhil Bhat, Graduate School of Business Columbia University. Vivek F. Farias, Sloan School of Management Massachusetts Institute of Technology. Camac C. Moallemi, Graduate School of Business Columbia University. Deeksha Sinha Operations Research Center Massachusetts Institute of Technology. https://web.mit.edu/vivekf/www/papers/experiment_design.pdf

[4] Rasmus Ros rasmus.ros@cs.lth.se Lund University Sweden. Per Runeson per.runeson@cs.lth.se Lund University Sweden. Continuous Experimentation and A/B Testing: A Mapping Study. 2018 ACM/IEEE 4th International Workshop on Rapid Continuous Software Engineering. https://ieeexplore.ieee.org/document/8452106

[5] Ron Kohavi1 and Roger Longbotham2 1Application Services Group, Microsoft, Bellevue, WA, USA 2Microsoft, Data and Decision Sciences Group, Redmond, WA, USA. Online Controlled Experiments and A/B Testing. DOI: 10.1007/978-1-4899-7687-1_891. https://www.researchgate.net/publication/316116834_Online_Controlled_Experiments_and_AB_Testing.

[6] Frank P.A. Coolen. Dept. of Mathematical Sciences, Durham University, Durham, United Kingdom frank.coolen@durham.ac.uk. Some statistical aspects of software testing and reliability. https://maths.durham.ac.uk/stats/people/fc/Springer12-preprint2.pdf.

[7] Coolen FPA, Utkin LV (2011) Imprecise reliability. In: International Encyclopedia of Statistical Science; Lovric (Ed.). Springer, Berlin, 649–650. DOI: 10.1007/978-3-642-30662-4_8.

[8] Andrea Arcuri and Lionel Briand Simula Research Laboratory, P.O. Box 134, Lysaker, Norway. Email: {arcuri,briand}@simula.no. A Hitchhiker’s Guide to Statistical Tests for Assessing Randomized Algorithms in Software Engineering. Technical Report, Simula Research Laboratory, number 2011-13. https://www.simula.no/sites/default/files/publications/Simula.simula.670.pdf.

[9] Giordano Tamburrelli and Alessandro Margara Faculty of Informatics. University of Lugano, Switzerland. {giordano. tamburrelli | alessandro.margara}@usi.ch Towards Automated A/B Testing. https://www.researchgate.net/publication/264435539

[10] Shafi Kamalbasha and Manuel J. A. Eugster Avira Operations GmBH & Co. KG Tettnang, Germany. Bayesian A/B Testing for Business Decisions. https://doi.org/10.48550/arXiv.2003.02769.

[11] Maria Esteller-Cucala, Universitat Politècnica de Catalunya-Barcelona Tech., Barcelona, Spain. Vicenc Fernandez, SEAT, S.A., Barcelona, Spain. Diego Villuendas TechTalent-Lab, Department of Management, Universitat Politècnica de Catalunya-Barcelona Tech., Barcelona, Spain. The AB Testing Pitfalls Companies Might Not Be Aware of—A Spotlight on the Automotive Sector Websites. https://doi.org/10.3389/frai.2020.00020

[12] Komang Candra Brata, Adam Hendra Brata Mobile, Game, and Media Research Group, Department of Computer Science, Universitas Brawijaya, Indonesia. User experience improvement of Japanese language mobile learning application through the mental model and A/B testing. DOI: 10.11591/ijece.v10i3.pp2659-2667

[13] Isak Kabir, How to conduct A/B Testing? https://towardsdatascience.com/how-to-conduct-a-b-testing-3076074a8458

[14] [A. Gelman, J. Carlin, H. Stern, D. Dunson, A. Vehtari, and D. Rubin, Bayesian Data Analysis, 3rd ed. Chapman and Hall/CRC, 2013

[15] J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[16] I. S. Jacobs and C. P. Bean, “Fine particles, thin films and exchange anisotropy,” in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

[17] K. Elissa, “Title of paper if known,” unpublished.

[18] R. Nicole, “Title of paper with only first word capitalized,” J. Name Stand. Abbrev., in press.

[19] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, “Electron spectroscopy studies on magneto-optical media and plastic substrate interface,” IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[20] M. Young, The Technical Writer’s Handbook. Mill Valley, CA: University Science, 1989.