The Q-q Dynamic for Scientific Education[1]

Xiao-Li Meng

Department of Statistics, Harvard University

1. Quantify Qualification and Qualify Quantification

I surmise that most readers would agree that a dialogue of qualitative thinking versus quantitative thinking is unlikely to be fruitful, or even meaningful, without defining what these two thinking processes entail. However, the very desire for this definition highlights the difference between, and the need for, the two processes. For those who have strong training in quantitative thinking, the phrase "defining" may induce a thought process to specify the boundary of each to the exclusion of the other, before attempting to argue their values in a contextualized framework. Those who feel more at home with qualitative thinking may rely on context to prompt readers' own interpretations, or rather contextual associations, to discern between the two, as the dialogue unfolds.

Indeed, any reader who tries to locate the precise definitions of these two types of thinking in the article by Ograjensek and Gal (2014, hereafter OG) would be disappointed. A minor dose of qualitative thinking of itself, that is, the impossibility of quantifying qualitative thinking, would help such readers to avoid the subsequent mistake of ignoring OG's key messages. OG (Section 2.1) apparently associates quantitative thinking strongly with "skills and methods" taught in statistical courses, and qualitative thinking with the need for "an external point of view, on the assumption that a learner can better understand the strengths but also the limitations of one system only by comparing to another." Furthermore, OG associates qualitative thinking as a skill set that is needed by "an overlooked, yet huge, group of statistics learners" who "do not aspire to become statisticians," with the apparent implication that such skill sets are not taught or taught well in our current "service courses."

I cannot agree more strongly with OG's emphasis on enhancing the qualitative thinking in our statistical education, regardless of the students' aspirations or career orientations. Indeed, without critical qualitative thinking, the field of statistics would largely lose its identity, as a colleague expressed succinctly: "qualitative thinking is what makes doing statistics different from just applying statistical software … I've found that many applied researchers (and intro stats teachers) are not aware of the distinction at all.'' And similar emphases are being further stressed across a number of evidenced-based disciplines. For example, a most recent article by Green et. al. (2014) investigated the critical role *mixed methods*, by which they mean integrating quantitative and quantitative methods, can play "to increase depth of understanding while improving reliability and validity of findings" in mental health studies.

Intriguingly, OG is itself a demonstration of the effectiveness of the qualitative approach, a style I appreciate especially as a discussant, because it has provided me with much food for thought. Evidently, readers will have their own experiences and understandings of what are taught, and not taught, in the statistical courses with which they are familiar, and it is in such specific contexts OG's messages can be

[1] To appear in *International Statistical Review* (Version June 1, 2014).

appreciated in practical terms and hence have the most impact. Of course, such an approach is not without risk, because some readers can (and will) dismiss OG's premise as a red herring, for everything described in OG as being missing or lacking emphasis might already have received much attention in these readers' own teaching and practice. Indeed, as OG quoted, many points they raised were already discussed and implemented by those who have been deeply concerned with statistical education and literacy.

Yet this is precisely the reason that OG's qualitative approach is working. If OG had attempted to quantify qualitative thinking or qualify quantitative thinking irrespective of readers' interpretations or associations, I am afraid OG's key message of ensuring the interplay of the two processes to have direct impact on statistical education would be marginalized. An article can only carry effective messages to those who can be effectively affected by it. It is clear that OG's goal was to raise the awareness of the importance of qualitative thinking in statistical education among those who have not given it adequate attention, and to provide much motivation for them to adopt suggested actions in nimble and operative ways. It is therefore fitting for OG to practice what it preaches by avoiding preaching to those who already practice.

2. The Interplay of Qualitative and Quantitative Thinking

Having done my own preaching, let me practice what I preached by mapping out what I consider as qualitative thinking versus quantitative thinking in three specific contexts. These examples are meant to supplement OG's general discussions and emphases, but I am mindful that readers' qualifications (or quantifications) may differ substantially from mine. The three classes of problems are chosen because they are at the center of the "Big Data" era, requiring extensive interplay of quantitative and qualitative thinking, and yet they are virtually absent from current textbooks. For a more detailed treatment, see Meng (2014).

2.1 Multi-resolution Inference

The discipline of Statistics is full of insights that require both qualitative and quantitative thinking in order to realize their full value. The ever-prevailing bias-variance trade-off, or more generally robust-efficiency trade-off, is a telling illustration. Realizing its ubiquitous nature, especially in situations where it is far from obvious, requires critical qualitative thinking; however, figuring out how to make the trade-off in specific settings calls for trained quantitative thinking. (And please take this as a qualitative definition, not a quantitative one!) Indeed, without the qualitative understanding that it is impossible to have a purely data-driven *automated bias-variance trade-off*, one could easily devote energy to chasing this quantitative phantom, as explained in Meng (2009) and Blitzstein and Meng (2010).

Consider the ever-increased attention to personalized treatments, such as personalized medicine, personalized education, etc. These all sound heavenly – who does not want a treatment that would be guaranteed to work for *me*, instead of for some average/representative person, however defined? But a bit of qualitative thinking would compel one to ask where on earth could anyone find enough data to establish that a treatment would actually work for *me*? A question then arises naturally: whenever someone claims to have a personalized treatment, what was the statistical and scientific evidence to support that claim? Raising such skepticism does not require one to understand all the nuts and bolts of causal inference (and indeed few could claim they do!), but simply a qualitative understanding of how

usual clinical trials are conducted – who would be the right guinea pigs for *me*? This is by no means to suggest that the search for personalized treatments is wrong-headed, but rather that a healthy dose of critical thinking is needed more than ever as we bring science and technology to a deeper level, literally and figuratively, and that typically such critical thinking is of a qualitative nature instead of a quantitative one.

Yet critical qualitative thinking is also essential for motivating enhancement and new directions in quantitative research, a point stressed by OG as well (Section 2.2). Accumulating statistical evidence for personalized treatment requires us to reverse the inference direction typically taught in current introductory courses, that is, our aim is no longer to "go up" from a sample to a population, but rather to "go down" from a sample or population to an individual. This realization compels us to re-think our usual asymptotic setup with n approaches infinite. We really don't have any n to speak of, not even $n=1$ (but see below), because trivially there cannot be any *direct* data on how a treatment works for me before I get treated by that treatment. Even if I had a "naturally and nurture-ly" identical twin brother who served as my guinea pig, I still would only have $n=1/2$, because my twin brother can only undergo one treatment, with the other being counterfactual.

This leads to the multi-resolution inference framework, where we explicitly acknowledge the resolution level, that is, granularity, at which we collect and analyze our evidence. Consider for example a scenario in which my doctor has data comparing two treatments among a population of people for whom we know gender, age, weight, height, and blood pressure. If the doctor does not think height would affect the treatment outcomes, she may just consider those who have the same gender, age, and similar weight and blood pressure as mine; that is, she is effectively forming a resolution 4 equivalence class for me. If height is suspected to be relevant, then she should consider stratifying on all 5 variables, and hence a smaller equivalence class of resolution 5. In general, the higher the resolution, the more relevant the equivalence class would be for "me," but at the expense of smaller sample size. It is therefore a classic case of bias-variance trade-off.

Of course, determining the optimal or even a reasonable trade-off will require much quantitative thinking, but the motivation comes from the qualitative thinking that the trade-off is what makes it possible to accumulate statistical evidence for personalized treatments. This realization also helps to build a natural bridge from the real world to the counterfactual world underlying much of the potential outcomes framework for causal inference (e.g., Imbens and Rubin, 2014). That is, the counterfactual world can be viewed as the limit of real world when the resolution level goes to infinite, where "me" becomes unique, and hence can only be assigned to a single treatment (at a given time). See Liu and Meng (2014) for further discussion, and an illustration of how the interplay between qualitative thinking and quantitative thinking carries out in the context of resolving the Simpson's paradox from the multi-resolution perspective.

Incidentally, during the preparation of this discussion, I learned that for certain chronic diseases, there has been an increasing interest in the so-called "n-of-1" trial, using a patient him/herself as the guinea pig. Qualitative thinking is still of critical importance here, to identify situations in which "me" is not automatically a good guinea pig for myself, because different treatments are necessarily applied over different time periods. The "me" in the warm and sunny week one might be a very different animal from the "me" in the cold, rainy, and windy week two! If the effect of time is deemed to be less important than

the cross-sectional factors (e.g., gender, race), an n-of-1 trial is indicated that places the higher priority on controlling for the cross-sectional factors; otherwise, a parallel group trial design might be more appropriate. Other confounding factors that need to be addressed include the ordering effect, the carryover effect, and the onset effect; see Duan et. al. (2013) and Kravitz et. al. (2014). Assessing the degree of bias caused by such confounding factors obviously will require quantitative methods, but these methods become relevant only after one realizes the presence of these confounding factors in the first place, an identification process that belongs squarely to qualitative thinking.

2.2. Multi-phase Inference

Another great example of this interplay is in the context of multi-phase inference, which originated from multiple imputation inference (Rubin, 1987) under uncongeniality (Meng, 1994). Any large-scale data sets available for analysis, especially those in the public domain, are never "raw data," however defined, but an output of a "data cleaning" or "data repair" process. Unfortunately, this fact is not always appreciated, as OG pointed out, "We find it unacceptable that not only students but also some experienced researchers tend to treat official (in fact any form of secondary) data as by default purely quantitative and not subject to measurement errors." Indeed, the "errors" here go way beyond the usual measurement kinds, because common cleaning and repair processes include parts or all of re-calibration, re-normalization, compression, outlier removing, imputation, etc. Furthermore, there is an entire---and rapidly growing---enterprise of protecting privacy by purposely distorting individual data points, while aiming to reasonably preserve their statistical distributions, both marginally and jointly (a good resource for relevant literatures and many open research questions is the open-access *Journal of Privacy and Confidentiality* hosted at repository.cmu.edu/jpc).

Regardless of the actual preprocessing, a bit of qualitative thinking would help us to realize that such a preprocess must have an impact on our final analysis, and that the impact can be both desirable and undesirable, depending on how the preprocessing is done. The impact part is easy to understand, and so is the desirable part---otherwise why bother with preprocessing at all? Deeper qualitative thinking is needed, however, to realize that undesirable impact is also inevitable, even when the original goal is well-intentioned and those individuals who carry out the preprocessing have done an absolutely perfect job given the resources and information available to them.

The inevitability of an undesirable impact reflects the sequential nature of the multi-phase inference, where the final inference conclusion is a result of efforts from multiple phases: data collection (pre- or post-study design), data cleaning, data analysis, etc. These phases have to be time ordered, and those who work on different phases typically do not---or even are not allowed to (e.g., due to data confidentiality)---have full knowledge of how other phases were or will be carried out. Consequently, the assumptions made at an earlier phase may not be comparable with that of later phases. This leads to the problem of uncongeniality (Meng 1994, Xie and Meng, 2014), which basically means that there is a consequential discrepancy between assumptions made at different phases, and the consequences can be far from desired.

A common example in the context of multiple imputation is that at the imputation stage a covariate, say, Z, was deemed to have little additional predictive power for imputing the missing response Y given all other covariates, say X, and hence Z was not included in the imputation model. However, some users of the imputed data may be only interested in the relationship between Y and Z. When such users analyze the multiple imputed data sets, even if they follow all the proper procedures (e.g., as given in Rubin 1987),

they will still have a biased estimator of this relationship because it is weaker in the imputed data than in reality. The deeper concern is that these users may not have any quantitative way to correct this bias because, *at best*, they are only aware of a qualitative description of the imputation model (e.g., consider how many users of, say, the US Census Bureau's datasets are actually aware of the Bureau's approaches to imputation, or have the interest and resources to find them out). Such qualitative knowledge is generally insufficient for deriving quantitative bias corrections.

No one intended negative impacts, but they are inevitable because no imputer can possibly anticipate all types of subsequent analyses of imputed data sets. Even if this were possible, there will not be enough data to build an encompassing and *saturated* imputation model to include all covariates of potential interest --- such a model will have little predictive precision because of overfitting. Imposing assumptions to effectively lower the dimensions (such as via LASSO) would reduce such overfitting, but then it essentially defeats the purposes of including all of the variables to make the imputation model as saturated as possible (Meng, 1994). Therefore, once again we face a bias-variance trade-off, but this time it is not even clear how to formulate the optimal trade-off because there are many users involved. It does not take much qualitative thinking to realize that what is optimal for one user is likely to be suboptimal for another when two users have different analysis objectives, and yet imputation is meant to be general-purpose, at least for public data files.

However, this by no means suggests that nothing can be done quantitatively. To the contrary, it has motivated and will continuously motivate quantitative research in new directions. As a matter of fact, it motivated me to formulate the concept of uncongeniality and to introduce the notion of Bayesian model embedding of a frequentist analysis procedure (Meng, 1994). This embedding makes it possible to quantify qualitative differences between a Bayesian imputation model and a frequentist analysis procedure, often given by statistical software. Much more research is needed to identify quantitative indices that can best capture the degree of uncongenialities in many common practical settings, as well as on how the ultimate biases are quantitatively related to such indices; see Xie and Meng (2014) for some initial explorations and results. More broadly, what should be the appropriate statistical theoretical foundation of multi-phase inference in general, and what are the quantitative results we can establish that will provide both theoretical insight and practical guidelines? Again, preliminary investigation and findings, as reported in Blocker and Meng (2013), suggest that for this type of research the ability *to think qualitatively but to act quantitatively* is of paramount importance. Much of the needed deep thinking was/is not about how to derive or prove mathematical results, but rather about how to build conceptual frameworks and understand practical constraints, which are typically hard to quantify but necessary for the theory of multi-phase inference to be relevant in practice.

2.3. Multi-source Inference

A common misconception of "Big Data" is that *big* implies more and hence better information. But Big Data also means Big Noise, and this realization itself only requires qualitative thinking – it is wishful thinking that more data only bring in more information, not noise. It is entirely possibly that more data mean worse results, if we do not know how to properly extract information from the data, and this phenomenon can and did happen with procedures as common as ordinary least squares, as reported in Meng and Xie (2013). A recent somewhat surprising experience regarding "Big Data" reinforces OG's emphasis that qualitative thinking is lacking, even (especially?) among some professional statisticians.

I was giving a talk at a statistical conference, and one part of my talk was about multi-source inference for "Big Data." A key feature of multi-source inference is that at least a part of the data were not collected at all for statistical inference purposes, and they tend to dominate in size, e.g., a national database on employment insurance. An intriguing question then is, for the purpose of statistical inference, how useful are such large-scale non-probabilistic datasets compared to a much smaller but probabilistic sample? I therefore asked the audience which one they would trust more: a 5% random sample or an 85% non-random sample?

A bit of qualitative thinking would go a long way here. Whereas the meaning of "random sample" is or should be clear to statisticians, "non-random sample" could mean anything, and clearly how much one can trust it would depend on how "non-random" it is. Therefore a sensible answer, without any further information, would be "It depends!" It could also be justifiable to prefer the 5% random sample, on the grounds that it will always deliver a valid answer, regardless of the quality of the 85% sample. (In real-life situations, one of course should also question whether any sample can be truly "random," that is, free of any defect that could induce biases.)

But interestingly, more than half of those who responded placed their trust in the 85% non-random sample, and when I asked for reasons, the response was "85% seems large enough." Since these respondents were all statisticians, it is very unlikely that they did not realize the potential bias caused by a non-random sample. Therefore the thinking must be "85% is much larger than 5% and is close enough to 100%, and hence it cannot go too wrong."

This is a good example of how quantitative thinking can mislead in the absence of accompanying qualitative thinking (as outlined above). Indeed 85% is much larger than 5%, but this quantitative comparison is meaningful only when the two samples being compared have the similar quality. Indeed, even if the non-random sample reaches 95% of the population, it could still be much worse than a 5% or even 0.05% random sample in terms of mean squared error (MSE) of the sample mean. Here a bit of quantitative thinking helps. For a random sample, the MSE is the same as the variance and hence it is controlled by the *absolute sample size n*, precisely by $(1-f)/n$, where $f=n/N$ and N is the population size. In contrast, for the non-random sample, the MSE is dominated by the square of the bias term when the sample size is large, and hence it is essentially controlled by the *relative sample size*, more precisely by $1-f$. Clearly $1-f$ goes to zero much slower than $(1-f)/n$ as n grows, and indeed $1-f$ can be made to stay away from zero while $(1-f)/n$ approaches zero arbitrarily closely, when N is very large.

An illustrative example was given in Meng (2014), which also introduced the notion of *data defect index* (ddi). For a sample mean, a good ddi is ρ, the correlation between the value of a data point and its chance of being observed/recorded in the non-random sample; hence, when $\rho=0$, the sample can be viewed effectively as a random sample (for estimating the population mean). Then for a non-random sample with $\rho = 0.1$, it will take more than f=95% sample to guarantee beating a random sample of n=2400 from the same population in terms of the MSE of the sample mean. Because this comparison is irrespective of the actual population size N, the percentage of the random sample n=2400 can be made arbitrarily close to zero as N grows. This example highlights the need to emphasize that whenever we think to rely on the "Big" in the "Big Data" to ignore a potential bias, we need to remind ourselves that it is the relative size $f=n/N$ that matters, not the absolute size n, and the larger the ddi, the larger f needs to be to achieve the same amount of bias reduction. We need to teach our students more about such qualitative thinking, so

they will not be mis-impressed and hence mislead by big numerical values of *n*. A biased sample with n=10,000,000 (and with a non-negligible ddi) can lead to a far worse inference than a simple random sample of 1,000 or even 100, if the population size is, for example, N=300,000,000.

3. A Q-q Dynamic to Codify the Interplay

OG's key point is that qualitative and quantitative thinking co-exist and interact at all research stages, and therefore there should be an on-going emphasis of this interplay in all statistical education and beyond. I cannot agree more. As previous examples demonstrate, the two thinking processes typically interweave and enhance each other. If any separation of the two processes is desirable for pedagogical purposes, it is merely a matter of their different degrees of emphasis at different stages. For example, qualitative thinking naturally is more dominating at the exploratory stage because the very meaning of exploration includes figuring out what should---and can---be quantified. However, researchers with little training in quantitative methods are less likely to complete a meaningful exploratory study on their own, because they are more likely to have difficulties to fully anticipate or even appreciate the complications of data collections, preprocessing, or analysis that come with every real-life study.

To help to highlight the critical importance of this on-going interplay, I suggest we codify it as a *Q-q dynamic*, with Q representing the thinking process receiving more emphasis at a particular stage. When and which "q"---quantitative or qualitative---deserves to be capitalized will depend on the context. Hence for the exploratory stage, Q more likely is reserved for the qualitative thinking, and in the analysis stage, Q may be assigned to the quantitative thinking more often. But in both stages the other q also presents, as OG emphasized, and hence we must constantly stress to our students the Q-q pairing and their complementary nature.

Whereas the Q-q dynamic is by no means restricted to statistical education, perhaps a useful (but of course not perfect) statistical analogy is to link qualitative thinking to non-parametric methods, loosely structured to allow for greater freedoms for explorations, and to link quantitative thinking to parametric methods, highly structured to induce more precise and focused summaries. I surmise most statisticians would agree that there is no intrinsic contradiction between non-parametric methods and parametric methods, because they serve different purposes, and their pairing is what makes it possible for us to move freely across the entire spectrum of robust-efficiency frontier. I am therefore in full agreement with OG's point of avoiding treating quantitative thinking and qualitative thinking as two separate schools of thoughts, but instead "to look not only for differences but also for *similarities* and for *common elements* when thinking about education of would-be statisticians (i.e., statistics majors) as well as basic statistics education as part of introductory or service courses in statistics."

I believe an effective means to achieve OG's pedagogical aims is to provide many examples and illustrations to ingrain the dynamic of the Q-q pairing into students' overall learning and thinking process. For example, we can emphasize that the realization of the existence of selection bias in statistical inference is qualitative thinking, whereas figuring out how much bias is there, or how to reduce it, is quantitative thinking. More generally, as OG alluded, the realization that any method comes with its own limitation is qualitative thinking, and to figure out what the limitations are requires more quantitative thinking. And finally, understanding the importance of quantitative thinking is qualitative thinking, but to assess the damage done by lack of or false qualitative thinking requires quantitative thinking.

The very first sentence of OG, "The scholarly dialogue on what constitutes statistical training for majors and non-specialists alike has been initiated several times in the past, yet never brought to a conclusion," is a fact. Any attempt to bring it to a "conclusion," however, would go against the very message OG tries hard to convey. The notion of "conclusion" reflects our, perhaps subconscious, desire for a quantifiable single answer, a static state of mind we hope can motivate and advance our future endeavors in the said course. However, such a desire/hope always exists, and therefore we may as well take full advantage of its self-perpetuating inertia, as long as our force of qualitative thinking is strong enough to direct and re-direct its trajectories indefinitely. That is, the real driving force for educational betterment must be the process itself, just as the Q-q dynamic is always a process, not an end on its own. OG rightly assigned Q to qualitative thinking for their general messages, and I hope our collective emphasis on the Q-q dynamic will lead to some real educational impact that we all are keen to witness. And the field of Statistics should serve as a leading light in scientific education, for which understanding and teaching the interplay of qualitative and quantitative thinking, that is, the Q-q dynamic, is essential.

Acknowledgments

My thanks go to Professors Irena Ograjenek and Iddo Gal for an inspiring article, to Professor Christopher Wild for inviting me as a discussant, and to Joe Blitzstein, Richard Cleary, Naihua Duan, Andrew Gelman, Cassandra Wolos Pattanayak, and Jeremy Wu for very helpful comments and exchanges. I also thank US National Science Foundation for partial financial support.

References:

Blitzstein, J. and Meng, X.-L. (2010). Nano-project qualifying exam process: An intensified dialogue between students and faculty. *The American Statistician*, **64,** 282-290.

Blocker, A. W. and Meng, X.-L. (2013). The potential and perils of preprocessing: Building new foundations. *Bernoulli*, **19,** 1176-1211.

Duan, N, Kravitz, R. L, and Schmid, C. H. (2013). Single-patient (n-of-1) trials: a pragmatic clinical decision methodology for patient-centered comparative effectiveness research. *Journal of Clinical Epidemiology*. **66** (8 Suppl):S21-8.

Green, C. A., Duan, N., Gibbons, R. D., Hoagwood, K. E., Palinkas , L. A. and Wisdom, J. P. (2014). Approaches to mixed methods dissemination and implementation research: methods, strengths, caveats, and opportunities. *Administration and Policy in Mental Health*. 2014 Apr 11. Available at *http://www.ncbi.nlm.nih.gov/pubmed/24722814*. (Epub ahead of Print.)

Imbens, G. and Rubin, D. B. (2014). *Causal Inference for Statistical, Social, and Biomedical Studies: An Introduction.* Cambridge University Press.

Liu, K. and Meng, X.-L. (2014). A fruitful resolution to Simpson's paradox via multi-resolution inference. *The American Statistician***, 68**, 17-29.

Kravitz, R. L., Duan, N, eds, and the DEcIDE Methods Center N-of-1 Guidance Panel (Duan, N., Eslick, I., Gabler, N. B., Kaplan, H. C., Kravitz, R. L., Larson, E. B., Pace, W. D., Schmid, C. H., Sim, I.,

Vohra, S,). (2014) *Design and Implementation of N-of-1 Trials: A User's Guide*. AHRQ Publication No. 13(14)-EHC122-EF. Rockville, MD: Agency for Healthcare Research and Quality. Available at www.effectivehealthcare.ahrq.gov/N-1-Trials.cfm.

Meng, X.-L. (1994). Multiple-imputation inferences with uncongenial sources of input (with discussion). *Statistical Science*, **9,** 538-558.

Meng, X.-L. (2009). Automated bias-variance trade-off: Intuitive inadmissibility or inadmissible intuition? In *Frontiers of Statistical Decision Making and Bayesian Analysis (Eds: M.C. Chen et. al.)*, 95-112. Springer.

Meng, X.-L. (2014). A trio of inference problems that could win you a Nobel prize in statistics (if you help fund it). In *Past, Present, and Future of Statistical Science (Eds X. Lin et. al.)*, 536-562. Available at http://www.stat.harvard.edu/Faculty_Content/meng/COPSS_50.pdf.

Meng, X.-L. and Xie, X. (2014). I got more data, my model is more refined, but my estimator is getting worse! Am I just dumb? *Econometric Reviews* **33**: 218-250.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons.

Xie, X. and Meng, X.-L. (2014). Dissecting multiple imputation from a multiphase inference perspective: What happens when there are three uncongenial models involved? Submitted to *Statistica Sinica.*