

## **“Name That Segment!”: Questioning the Unquestioned Authority of Numbers**

DONNA K. FLYNN  
TRACEY LOVEJOY  
*Microsoft Corporation*

DAVID SIEGEL  
SUSAN DRAY  
*Dray & Associates, Inc.*

*In many companies, numbers equal authority. Quantitative data is often viewed as more definitive than qualitative data, while its shortcomings are overlooked. Many of us have worked to marry quantitative with qualitative methods inside organizations to present a fuller view of the people for whom we develop. One area of research that increasingly needs to blend quantitative and qualitative methods is user segmentations. Our software technology product team has been using a segmentation based on quantitative data since 2005. One outcome of this effort has been the development of an algorithm-based “typing” tool intended to be used as a standard tool in recruiting for all segmentation-focused research. We learned that the algorithm was an indecipherable black box, its inner workings opaque even to those who owned it internally. This case study looks at how qualitative research came up against the impenetrable authority of a quantitative segmentation and its associated typing tool, and subsequently contributed to the redesign of future segmentation methodologies and the integration of qualitative research as a key component of segmentation creation.*

### **PROLOGUE: THE “NAME THAT SEGMENT” GAME**

*It was a sunny day in Seattle, and we were setting up the “Greenhouse” room in our design studio—named as much for its warm, windowed corner location as its role in hosting analysis and brainstorming sessions—in preparation for a carefully choreographed encounter with our market research partners. David and Susan were organizing photos of our field participants on the whiteboards and walls, while Tracey and Donna reviewed the stories of each participant that we would share with our partners. It was the end of a long week of collaborative analysis, and the Greenhouse walls were thick with field photos, quotes, segmentation screener results, and scribbled post-it notes. Even the warmed windows were wearing the rich data we had been wading through all week, with our wise participants peering down on us from their perches of individual truths.*

*Our goal for the work-session was to share with our market research partners the conundrum that our qualitative research findings of the segments was not meshing with their quantitatively-based segment profiles. The disconnect became quickly apparent to us in the field, and although we had engaged our market research*

## Redefining What Is Core

*partners in multiple conversations about it, we knew they were still not fully understanding the implications of this disconnect: that the segmentation was seriously flawed by not accurately representing significant behavioral patterns. It was time to engage our partners experientially, and let them bring their own interpretation to the qualitative data.*

*The three market researchers who collectively owned the segmentation—and with whom we had a longstanding, strong relationship—soon joined us in the Greenhouse. Six of us clustered around the table in the richly cluttered data room, while Tracey moved around on foot to point out 9 different participants as we recounted their stories. For each participant, we orally painted their portrait: “Joan is 41 years old, lives with her husband and daughter, and owns 2 small businesses, a sandwich shop and a flower shop. She is planning to buy a new phone as soon as her contract is up, likely a BlackBerry. It is important for her to try to always answer calls because ...” At the end of each story, we asked our partners to privately write down what segment they thought that person classified into, before sharing them openly and discussing...*

### CONFRONTING THE BLACK BOX: BEHIND THE SCENES OF A SEGMENTATION REFRESH

In 2005, Microsoft market researchers launched the Windows Mobile division’s first large-scale quantitative effort to understand and segment the mobile-phone owning population, aged 16-65, in the United States and Western Europe. Although there had been some prior efforts to create typologies for mobile technologies, this was the first time the business was provided a comprehensive and rigorous segmentation. The initial methodology consisted of quantitative surveys across priority markets and did not include any qualitative components, such as to inform the segmentation survey or validate the segmentation model.

One output from the segmentation was a typing tool that was used to recruit people into the segments for future research, using an abbreviated list of questions from the large survey. This is a necessary deliverable for any segmentation, and many researchers in both marketing and engineering—including ourselves—relied on it to recruit participants for various studies. Although we used it extensively, we didn’t really understand the algorithms or analytic assumptions that drove the tool or why the items were determined to be the most important ones. We had no transparency into the algorithm, and nor did we dig deeply into it of our own accord. In those early segmentation days, the typing tool had a fairly high accuracy and we became habituated to using it.

The absence of qualitative components in this first segmentation effort presented an opportunity for Donna’s UX Strategy team<sup>1</sup> to step in and examine some of the questions around the framework and accrue deep knowledge about these new segments that were

---

<sup>1</sup> In 2005 Donna Flynn started a Windows Mobile UX Strategy Team with the charter to drive target market and customer understanding throughout the development cycle from a user experience perspective, as a partner and complement to the work done in the marketing teams. The team consisted of three people: Donna Flynn, Rive Citron and Tracey Lovejoy, two of whom are authors.

## ***Redefining What Is Core***

virtually unknown to the business. The ethnographic research and subsequent deliverables around the segmentation grew to have broad impact, as gauged by organizational adoption of our customer tools (personas, scenarios, frameworks), the depth of target customer knowledge we instilled in our executive leadership and individual contributors across the division, and the influence of some of our models on a number of critical business decisions. On the other hand, our market research colleagues continued to dismiss the value of our ethnographic work on the segmentation itself, and saw it only as filling in “day in the life” details for persona development.

The segmentation filled such a huge void for the organization that it was rapidly and enthusiastically adopted at all levels; finally the organization had a common framework for talking about target customers. The risky underside of this broad and deep adoption of the model only became apparent to us later when we saw that broad assumptions about characteristics of the segments were carried over to subsequent iterations of the model. Work around this early phase segmentation also set the stage for the rest of our story here – namely the way in which the statistical algorithms behind the segmentation became imbued with unassailable authority and created a veiled ‘black box’ around the analytics, until our qualitative work uncovered some severe flaws in the methodological model and contributed to its undoing.

As the business grew, the target markets increased and two additional regional segmentations were created. It was quickly acknowledged that this approach would lead to a proliferation of segments. Therefore a decision was made to combine the segmentations. The market researchers stewarded the analytics of the worldwide model with a vendor, but the larger stakeholder team that we belonged to had no insight into the process they pursued for re-factoring and defining the segments. We were at that time blissfully unaware that an intentional change was being made to the segmentation model—the cluster analysis itself was being based only on attitudinal questions regarding purchase drivers and all usage variables were removed from the analytics. This would have enormous implications moving forward.

The combined segmentation appeared, on the outside, to be an iterative evolution of the prior model—sharing all the same segments with one additional segment that heavily represented customer patterns evident in the Asian markets. This is important because it implied little change in the characteristics and differentiators of those segments that carried the same name as in the past model. The segments that everyone from executives to researchers to engineers had come to know so intimately were still there, seemingly unchanged. Consequently the broad assumptions about who comprised a segment and key differentiators for them remained intact.

## THE UNRAVELING OF A SEGMENTATION

In 2008 this ‘worldwide’ segmentation was refreshed, updating measurements of the sizes of the segments and adding five additional countries, bringing the total number of countries represented to twelve. The vendor proclaimed that although the size of the segments had changed, the segments themselves were still relevant. This refresh—again based only on quantitative data—presented another opportunity for the UX Strategy team to go back into the field. Since most of the segments persisted from the prior model, our approach was to update knowledge of what the segments were up to in real life in 2008—what had changed in the two years since we’d spent time with them? The study was explicitly not about validating the segmentation itself, since its apparent consistency with the prior segmentation imbued it with high internal credibility in the organization. Dray & Associates were hired to implement the field work, which combined lengthy in-home interviews with examination of artifacts, experience collaging, and day-in-the life timelines—with 43 participants across 2 priority markets, recruited using the updated segmentation algorithm-based typing tool.

### Finding Our Segments in the Field

Within days in the field, we knew something was terribly wrong—at least half of our participants did not fit into the segments as we had understood them. Because the basic behaviors and attitudes of the segments were supposed to have been consistent with the prior segmentation we did not understand why our recruited respondents did not match the segment definitions. This immediately raised questions at a number of levels. Was the algorithm behind this new screening tool fundamentally flawed? Had the segments themselves changed that much in two years? Had the screener been properly administered? Were these people outliers? Had people given misleading answers on the screener that did not reflect their actual purchase drivers?

We made some changes on the fly to our protocol to try to unpack these questions. We started to have our recruiter screen people according to both the old and new screening tools. We also took time with people during the sessions to re-administer the screener and explore discrepancies. Sometimes as a result of these discussions we identified how the respondents would have answered the questions had they interpreted the questions differently. We also began experimenting with the typing tool, to see how large or small changes in answers to individual items would change not only the resulting segment assignments, but the entire profile of scores for each person across all the segments.

Meanwhile, as the data was coming in, we maintained communication with our market research team to keep them apprised of our issues. Early on they began to acknowledge that there might be a difference between the new segmentation model based on purchase drivers and the old one based on purchase drivers and behavioral patterns, but they remained insistent that the former should still be meaningful for all aspects of product development.

## **Redefining What Is Core**

### **The Analytic Unveiling**

Re-administering the screening tool in the field gave us a chance to do an informal check on the reliability of the tool, i.e., the degree to which people's segment assignments were influenced by random factors.<sup>2</sup> We saw that people changed their answers frequently. Some of these changes were small, clearly attributable to simple inconsistencies in how people answer on a 7-point scale from one time to the next. Others were more dramatic, attributable to inconsistencies in the way people interpreted questions or specific words they found ambiguous from one time to the next. With ambiguous questions, an element of chance influences how people may interpret the question. What was particularly interesting was that even small changes in response sometimes produced changes in segment assignment. This indicated that segment assignment was quite sensitive to random fluctuations in responses on the screener. We later learned that the reliability of the typing tool had never been evaluated. Nor had it ever previously been validated<sup>3</sup> by demonstrating that it accurately predicted some important behavior in a new sample. Of course, our discovery that many people in our study did not fit the existing segment descriptions raised very strong questions about the tool's validity.

We also were concerned about the effect of the fact that the typing tool assigned people to whatever segment they received the highest score for. It was very rare for people to have a single segment score that was dramatically higher than the other segment scores. However,

---

<sup>2</sup> Reliability is defined conceptually as the opposite of noise in the data. High reliability says that there is, in fact, some phenomenon that is being measured and that scores are not random. In practice, reliability is demonstrated by showing that there is relative consistency across administrations of a measure at different times with the same people, or across different judges scoring the measure independently. For a measure that consists of multiple items trying to measure the same variable, reliability can also be demonstrated by showing that, if you divide the items in half randomly, scores on one half predict scores on the other half.

<sup>3</sup> While high reliability merely says that you are measuring something other than noise, demonstrating validity requires evidence that you are measuring what you think you are measuring, such as by showing that the measure predicts other theoretically related behaviors. Validity can be no greater than reliability, because only the non-random component of the scores can be valid. In addition to what appeared to be random changes in people's answers when we re-administered the segmentation tool questions, we also identified some systematic biases in how people interpreted certain concepts or words. Systematic biases create a problem with validity, as opposed to reliability. For example, some items involved the word "productivity." There was a tendency for people who used their mobile phones for work-related tasks to interpret this as being related to work. People who used their phones only for personal use often interpreted productivity much more broadly. For example, they might consider that a mobile phone made them more productive because they could make calls while in the car running errands. Another issue was that people who were less technically sophisticated tended to misinterpret technical terms in ways that made them score into more technically sophisticated segments. These people tended to think, for example, that all cell phones are by definition Wifi devices, and so had indicated that Wifi was extremely important to them, when in reality they did not use it at all.

## **Redefining What Is Core**

the algorithm would assign to Segment A, with equal apparent certainty, a person with a true peak on Segment A and a person whose Segment A score was only minutely higher than his Segment D score. If one had to bet on a single segment for any individual, the best choice would, of course, be to bet on the one with the highest probability. However, since the probabilities were spread fairly evenly across eight segments, people were actually more likely to belong to one of the seven segments other than their top-scoring one<sup>4</sup>. We found this absolute assignment to be haphazard and only sometimes aligned with people's true attitudes and behaviors.

Taking these factors together, we were able to explain some of the reasons why the typing tool was breaking down in identifying appropriate matches to the segments. The probing that we were able to do in the qualitative research provided us insights into some of the issues with the typing tool and allowed us to paint detailed pictures of the heterogeneity within each segment.

### **Penetrating the 'Black Box'**

Out of the field and back at Microsoft, we had several face-to-face meetings with the market researchers and finally came to understand two pivotal things that had previously been masked. First, the segmentation had changed dramatically because the clusters no longer took behavioral patterns into account, only mobile phone attitudes and purchase criteria. Secondly, the market researchers themselves were uncomfortable with the typing tool unequivocally assigning a person to a single segment because there was a lot of overlap among segments.

To the first point, even though the market researchers had made a conscious choice to not include behavioral patterns in the cluster analysis, they did not make this clear anywhere. Segment profiles that they published included descriptions of demographics, attitudes, *and* usage—but nowhere noted that survey items asking about behavior had not been entered into the analytics of clustering. In our ongoing discussions with our market research colleagues, they appeared to assume that purchase driver attitudes would predict actual behavior. In any case, to those of us accustomed to the previous segmentation, the segments were still very much the same—same name, same key characteristics, just new and growing sizes. This was completely misleading.

We also began to inquire in more detail about the inner workings of the algorithm for the typing tool. One thing that became apparent was that the decision rules incorporated into the algorithm were completely opaque. The vendor who developed the segmentation and the tool eventually acknowledged that the algorithm was so complex, with so many

---

<sup>4</sup> Imagine that you are betting on a card drawn at random from a deck that has one extra King. You would be wisest to bet that the card will be a King, even though there is a much higher probability it will be something other than a King.

## Redefining What Is Core

“trigger points” where differences in patterns of responses would kick someone into a different segment, that it could not be summarized intelligibly.

After we fully realized the extent to which things had changed inside the ‘black box,’ and consistently challenged the market researchers on the implications of this change for the business, their language around the typing tool began to also subtly change. They acknowledged that there was overlap among segments and that some people were ‘core’ to a segment, while others were more on the fringes. We were told that we should look for people that have ONLY a high probability in one segment, and not to recruit people who had probabilities bordering on more than one segment. Unfortunately, based on our recruiting efforts, this appeared to be a minute percentage of the population.

Additionally, our market research colleagues’ language about what being placed into a segment actually meant began to change. The determining questions for placing a person in a segment were about the importance of factors in choosing their *current* mobile phones. However, our qualitative research showed a number of cases where participants’ current phones did not even have the functions they rated as highly important. When we pointed this discrepancy out to our colleagues, they argued that these people were thus answering the questions aspirationally and therefore still mapped to that segment (with the underlying assumption that they would likely move into that segment with the next purchase). We disagreed, and argued back that having some people answer the questions based on current purchase drivers and others based on future purchase drivers inherently skews the data. It seemed to represent an effort to hang onto the idea that purchase driver questions would in some way predict behavior. Our protests were met with shrugs. So while the black box remained impenetrable, insights gained by our qualitative research began to make evident very clear ways in which the current model provided skewed data.

## EPILOGUE: PLAYING TOGETHER

*[Reprise] The three market researchers who collectively owned the segmentation—with whom we had a longstanding, collegial relationship—soon joined us in the Greenhouse. Six of us clustered around the table in the richly cluttered data room, while Tracey moved around on foot to point out photos of 9 different participants as we recounted their stories. For each participant, we orally painted their portrait: “Joan is 41 years old, lives with her husband and daughter, and owns 2 small businesses: a sandwich shop and a flower shop. She is planning to buy a new phone as soon as her contract is up, likely a BlackBerry. It is important for her to try to always answer calls because ...” At the end of each story, we asked our partners to privately write down what segment they thought that person classified into, before sharing them openly and discussing.*

*As we had anticipated, their answers were incorrect far more frequently than they were correct. Several times, each researcher had a different estimation of the appropriate segment for a participant. But often, they all agreed on what segment the participant seemed to represent. In general, they were quite confident in their guesses. In the end, our colleagues assigned a wide range of different segments to the 9 participants. Then we pulled the surprise out of our hat: all 9 of these people had been classified into the same segment by the*

## Redefining What Is Core

*algorithm-based typing tool. A new light of understanding began to shine in their eyes. By the end of the game, we had reached a shared understanding of the challenge that lay ahead of us all: how to come to terms with the fact that this segmentation that the organization had invested in and bought into didn't accurately classify people according to its assumed lines of differentiation.*

### GETTING INSIDE THE BOX: THE NEXT EVOLUTION OF OUR SEGMENTATION

Our story doesn't have an ending, yet. Partly as an outcome of what we share here, the segmentation model in question was fully retired and we launched a highly collaborative effort with our market research colleagues in building and defining a new segmentation. This is still in process, and we are optimistic that within a few months we will have a happy ending to our story. We would like to close by sharing with you some of the lessons we have learned and are now putting into practice in evolving our business' segmentation model:

*Get inside the black box, and know thy segmentation factoring.* During the update of our segmentation we blindly trusted the process and did not take the time to properly educate ourselves or ask probing questions about procedures. Even though our key competency is qualitative research, we found that it was essential for us to have basic knowledge of segmentation methodologies. Therefore the primary lesson for us is that we must take an active role in understanding segmentation methods and processes. You may have to educate yourself in fundamentals—as we have had to—or hire someone to work with who can deeply explain to you the pros and cons, and not just sell you the solution. In addition, be sure to ask questions along the way and do not settle for ambiguous answers or deferral to the vendor. The people in charge of the segmentation should be expected to be able to explain the details.

*Be part of the core working group.* In addition, when the segmentation update occurred we were not part of the small team that owned the decision-making process on how to approach the update or how to communicate the outcome. Thus, a key learning for us has been to ensure participation in the core segmentation working group. Most segmentations are created by a team, not by a single individual, and industry-wide it is increasingly accepted that the most successful segmentations are created by a group of people from different disciplines and teams (Bortner 2008). If you are having difficulty gaining access, two tactics that we have found to be compelling are carving out some official accountability for segmentation deliverables and/or offering money to co-fund the segmentation. We have established a stronger position at the table by owning and funding components of the next segmentation. If this level of participation is not possible, you could become a reviewer, stakeholder, or extended team member of the work.

*Incorporate behavioral factors in the mix of potential determinants.* The removal of behavioral patterns as a factor to determine the updated segments, in the end, caused a breakdown of the validity of the segmentation. As behavioralists, it did not occur to us that our colleagues



## **Redefining What Is Core**

would even consider removing behavior, and it has been surprising to us that some current segmentation guidelines still advocate for attitudinal variables over behavioral variables, going as far as calling behavior “backward-looking” and attitudinal “in many ways, perfect” (Bortner 2008). So the key lesson for us has not been the importance of behavior as a segmentation determinant, but to ensure that behavioral factors are included as potential segment determinants (especially if you are engaged in product or service development). However, we do not believe that behavior should be the only determinant, perhaps not even the predominant. The strongest segmentations are often based on multi-dimensional factoring of attitudes, behaviors, and aspirations/values.

*Segmentation definition should include quantitative and qualitative inputs.* Our original segmentation was defined solely with quantitative data, and our qualitative research was executed as a follow-up study to ‘better understand’ rather than to ‘help define’ or ‘validate.’ However, it became clear that our earlier work had been a key component in the success of the original segmentation and also provided critical insights to understanding why the updated segmentation was not accurate. We have learned to ensure that qualitative research is embedded as a component of segmentation definition. Qualitative research should ideally be included in both the front-end and back-end of the quantitative research. Qualitative outputs should directly influence development of quantitative tools, factoring decisions, and clustering decisions, such as to understand interpretations of questions and the range of answer choices to test the survey; to validate quantitatively defined segment clusters; to test the typing tool screener; to help define cluster interpretations; to help decipher which of the zillion data points are significant in distinguishing segments; and to gain a deeper understanding of the who, how & why of the segments.

We have learned valuable lessons across this segmentation journey—most importantly about our own accountabilities around ensuring more rigorous integration of qualitative and quantitative methods in the process. We have long created impact in our organization by representing the ‘voice of our users’ through rich storytelling and model development but learned that our success at positioning ourselves as storytellers can also limit our authority in contexts where numbers are valued above all. Through this journey we were able to break through some of these boundaries, and have forged a deeper working relationship with our market research colleagues, gaining mutual respect for our respective training, perspectives, and contributions in building a strong segmentation model for the business.

## **NOTES**

Most importantly we would like to thank our market research colleagues for staying in the game with us, as we continue to practice and learn new ways of playing better together. We’d also like to thank Brian Rink and Rick Robinson for their thoughtful reviews of drafts of this paper. The opinions represented in this paper are solely those of its authors and in no way represent an official position of Microsoft Corporation or any of its teams of market researchers, ethnographers, or user researchers.

**REFERENCES**

Berrigan, J.A., & Finkbeiner, C.T.  
1992        *Segmentation Marketing: New Methods for Capturing Business Markets*. New York, NY: Harper Collins.

Bortner, Brad with Ellen Daley, Heidi Lo  
2008        **Why Good Segmentations Fail: Five Steps To Have Segmentations Drive Business - And Your - Success**. Forrester Research.

Cooper, A.  
2007        *About Face 3: The Essentials of Interaction Design*. Indianapolis, Indiana: Wiley Publishing, Inc..

Cooper, A.  
2004        **The Inmates Are Running the Asylum: Why High Tech Products Drive Us Crazy and How to Restore the Sanity (2<sup>nd</sup> Edition)**, Pearson Higher Education.

Demumieux , R., & Losquin, P.  
2005        Gather customer's real usage on mobile phones. In *Proceedings of the 7th International Conference on Human-Computer Interaction with Mobile Devices & Services (MobileHCI)*, September 19-22, 2005, Salzburg, Austria. New York, NY: ACM. [doi>10.1145/1085777.1085828]

Ge, R., Ester, M., Gao, B. J. Gao, Hu, Z., Binay, B., & Boaz, B-M.  
2008        Joint cluster analysis of attribute data and relationship data: The connected k-center problem, algorithms and applications, July, *ACM Transactions on Knowledge Discovery from Data (TKDD)* , Volume 2, Issue 2.

Gibson, Lawrence D.  
2001        Is something rotten in segmentation?: What's right, wrong, and downright rotten with segmentation. *Marketing Research*, Spring. Pages 21-25.

Gownder, J. P. with Michelle de Lussanet and Dan Wilkos  
2009        **The Consumer Product Strategist's Guide To Segmentation Analysis: Don't Leave Segmentation To The Market Research Department Alone**. Forrester Research.

Jeon, M. H., Na, D. Y., Ahn, J.H. & Hong, J. Y.  
2008        User segmentation & UI optimization through mobile phone log analysis. In *Proceedings of the 10th International Conference on Human- Computer Interaction with*

## **Redefining What Is Core**

*Mobile Devices and Services (MobileHCI)*, September 2 – 5, 2008, Amsterdam, The Netherlands. Pages 495-496. New York, NY: ACM.

Kleinberg, J., Papadimitriou, C., & Raghavan, P.  
2004      Segmentation problems, March 2004, *Journal of the ACM (JACM)*, Volume 51 Issue 2, Pages 263- 280.

Sicilia, M-A., & García, E.  
2003      On fuzziness in relationship value segmentation: applications to personalized e-commerce, June, *ACM SIGecom Exchanges* , Volume 4 Issue 2, Pages 1 – 10.

Weinstein, A.  
1994      Market segmentation: using demographics, psychographics, and other niche marketing techniques to predict and model customer behavior. Chicago, IL.: Probus.