The Mystique of Numbers: Belief in Quantitative Approaches to Segmentation and Persona Development

David A. Siegel, Ph.D
Dray & Associates, Inc.
2007 Kenwood Parkway
Minneapolis, MN 55405 USA
david.siegel@dray.com

CHI 2010, April 10 - 15, 2010, Atlanta, GA, USA
ACM 978-1-60558-930-5/10/04.

Abstract
Quantitative market research and qualitative user-centered design research have long had an uneasy and complex relationship. A trend toward increasingly complex statistical segmentations and associated personas will once again increase the urgency of addressing paradigm differences to allow the two disciplines to collaborate effectively.

We present an instructive case in which qualitative field research helped contribute to abandoning a "state of the art" quantitative user segmentation that was used in an attempt to unify both marketing and user experience planning around a shared model of users. This case exposes risks in quantitative segmentation research, common fallacies in the evolving practice of segmentation and use of personas, and the dangers of excessive deference to quantitative research generally.

Keywords
User research, personas, segmentation, quantitative research, qualitative research

ACM Classification Keywords
D.2.2. Software engineering: Design tools and techniques; D.2.10 Software engineering: Design methodologies

General Terms
Human factors

Introduction

The paradigm differences between conventional market research and user-centered design (UCD) research are well known. There are many examples of situations in which the two approaches can lead to differing perspectives on customers for and users of technology. Sometimes these are complementary, but often they are contradictory.

The growth of segmentation and persona research will increase the urgency of integrating market research and UCD research. This case study focuses on challenges raised by two interconnected trends:

- Use of quantitative methods to develop a single user segmentation, and its representation in personas, to unite all product planning, design, and marketing around a shared model of the universe of users

- Using these segments and personas as the basis for screening user research participants

In many large companies, only marketing has the power and resources to drive a large-scale strategic segmentation and persona-development project. This means that traditional market research approaches tend to dominate the effort. We are specifically concerned about segmentations based on large N survey research. This paper presents a case in which a flawed quantitative segmentation had achieved a high level of premature buy-in partly because of the mystique of numbers. It raises important lessons about the limitations of quantitative segmentation research and the dangers of excessive deference to quantitative methods generally.

Some of the problems with the segmentation that we describe in this case study may seem obvious to those UCD researchers who are generally skeptical of market research, especially surveys. However, we have seen many cases in which UCD people have been swept up in the enthusiasm about quantitative segmentation research, which they see as providing a scientific basis for screening future research participants, without questioning whether the methodology really supports the claims. Also, when they are indeed uncomfortable with the quantitative approach, UCD researchers without a background in quantitative research may not know how to effectively critique the quantitative methodology itself. Both of these observations point to the need for qualitative UCD researchers to become more sophisticated in challenging the mystique of quantitative research.

Can Marketing and Design Use the Same Segmentation and Personas?

The goal of a consolidated segmentation that will make sense for marketing and for design is itself controversial. Alan Cooper, promoter of the use of personas in design, views marketing and UCD segmentations as fundamentally incompatible [2]. He argues that design requires distinctions among users related to likely usage, while marketing makes

distinctions based on demographics and focused on purchase decisions.

In support of the goal of integration, one can argue that, at a deep level, market research and UCD do have shared interests. Ideally, the marketing messages that will be compelling for a particular user segment should match the value actually delivered by the designed user experience for that segment. This means that both marketing and UCD need to focus on the match among audience, value propositions, and experience.

Also, it is a bit of a caricature to say that marketing is only interested in demographics, or that demographics have nothing to do with behavior. Demographics may simply be a crude proxy for things that are harder to measure on a mass scale, and certainly can have some correlation with behaviors of interest, even if they are only weakly predictive. Marketing is also interested in the "resonance" of its messages with users. Therefore, it frequently strives to go beyond demographics to try to understand the psychology of the purchaser. This can lead it into territory that overlaps with UCD.

However, even if the goal of creating a single unifying segmentation is laudable, there remain important differences in practice traditions between market research and UCD that have to be overcome. In particular, market research places much greater emphasis on surveys and tends to have more confidence that self-report predicts behavior (or worries less about the possible discrepancy). Many of the problems we uncovered in this case are directly attributable to those tendencies.

Project Background

Our client's market research organization had invested heavily over several years in large scale quantitative research to develop a user segmentation model for mobile phone users to organize efforts across all levels within the product group. The raw data was in the form of responses to rating scales regarding the importance to the respondent of various uses, value propositions, features, and phone characteristics in influencing their choice of phone. Based on statistical patterns in these responses from many thousands of respondents, they had identified 8 clusters, or segments, of respondents. Next, taking the meaning of the questionnaire responses that defined each segment at face value, they wrote descriptive profiles of each segment that included descriptions of how they used their phones. These became the initial personas, which were shared with the product teams.

An additional outcome of these efforts was a tool intended to screen and classify participants for future user research. The tool used a greatly abbreviated list of attitudinal questions from the larger survey together with a complex algorithm to assign new respondents to a segment. For each respondent, it calculated 8 scores, reflecting the degree of similarity between the respondent's answers and the response profile of that segment. It then assigned respondents to the segment for which they received the highest score. The tool based on this algorithm was implemented in Excel®--plugging the respondent's answers into a form yielded the numerical scores for each segment and identified the segment receiving the highest score.

Development of algorithms like these requires a great deal of computing power. They are not based on

simple correlations among the questionnaire answers and segment profiles, but on iterative testing of huge numbers of decision rules. They are tweaked until they do an optimal job of fitting the existing data set. The resulting decision rules are so complex that they cannot be summarized simply. This means that, in a sense, no one knows how the algorithm really works or what the real criteria are for segment membership. The algorithm itself becomes the operational definition of the segments.

By the time of our project, two major iterations of this quantitative research had been done. The more recent iteration extended the segmentation to some new countries. It identified the same number of segments as the earlier study. The researchers felt they could map the new segments onto the old ones, and so used the same segment names. This created the impression that the results confirmed the earlier findings, adding to the credibility of the segmentation.

One important change was made between the first and second iteration. In the first, along with the attitudinal questions, there were also frequency ratings for particular reported usage behaviors. In the second study, the latter were excluded from both the analysis and from the segmentation tool. As reported to us by our internal contacts, this was done on the basis of the belief that attitudes are "stable" but behaviors readily change.

By the time of our involvement, the segmentation and its associated tool had already gained a great deal of acceptance in the organization. The algorithm was claimed to have "80% accuracy." Also, we believe that the credibility of the segmentation was further enhanced by the fact that the massive amount of data that went into it, the sophisticated analysis, and the seemingly magical but opaque working of the algorithm all made this approach seem extremely "scientific." We suspect these factors also made it too intimidating to critique.

Our Project

We were engaged to gather rich qualitative data to support persona development for 4 high-priority segments out of the total of 8. Our mandate was We specifically not to validate the segments. Rather, we were expected to take them as a given. We spent 3 hours in homes with each of 43 participants in 2 countries, using in-depth interviews enriched by examination of people's phones for evidence of usage patterns, as well as by some projective exercises designed to elicit additional usage data.

The most recent segmentation tool formed the core of our screener. This was its first use in recruiting for in-depth, small sample research. Marketing's claim that the tool was 80% accurate led our clients to expect that it would classify people in our sample into fairly homogenous groups, making it a reasonable task to come up with a coherent persona to represent each of them.

Results

Almost as soon as we began our data collection in the field, we found that many individuals were dramatically different from the descriptive portraits of the user segments they supposedly belonged to, both in the meaning their phones held for them and in their behavior patterns,. The "official" segments in our sample seemed heterogeneous on just about every key

dimension. We found that we could easily group people into 2-4 subgroups within segments. Later, multiple iterations of a number of complementary qualitative analyses, carried out both blind and not blind to the "official" segment assignments, and using multiple judges, showed a high degree of consistency in these subgroup assignments.

What was even more striking was that all but one of the subgroups within a particular segment had a high degree of similarity to a corresponding subgroup in one or more different segments. We could combine similar subgroups together across segments (both in phone-related attitudes and behaviors) in ways that produced more coherent groups than when we grouped them by their assigned segments. In fact, the groups we created were essentially orthogonal to the official segments. Only one subgroup was unique to a segment, and no segment was made up mostly of people unique to it. Figure 1 shows the "official" segments and our qualitatively-defined subgroups, coded by number to show the matching subgroups across segments.

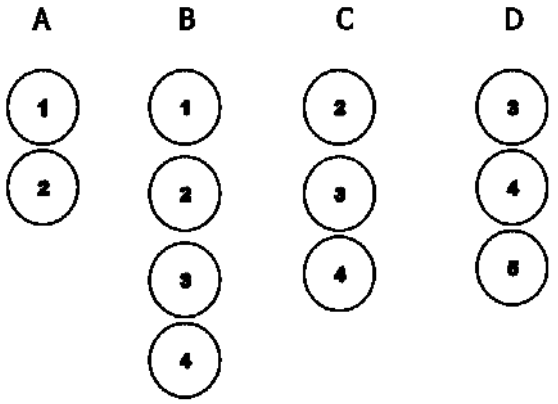

figure 1. Similar subgroups within 4 segments. Subgroup 5 was the only one unique to a segment.

How Can We Explain these Results?

Because our findings were so discrepant from expectations, despite the confidence that many at our client had in the segmentation tool, we had to be prepared for serious objections. We had to be prepared to show that our results were not an aberration, due perhaps to our small sample, or to some flaw in the recruiting. Most practitioners of UCD research would not be surprised that a segmentation based only on survey responses would do a bad job of predicting behavior, especially when it did not ask about behavior. However this argument was not likely to be convincing to people who were already believers in survey research, or to the people in the organization who were already convinced about the segmentation.

A more specific criticism that might explain our findings was that this segmentation did not factor behavioral variables into the definition of the segments--neither in the form of survey questions directly addressing behavior nor in the form of observational contextual research. Instead, it focused on attitudes that supposedly influenced purchase decisions. This may certainly be part of the explanation, but the groups in our sample created by the segmentation tool were heterogeneous in their attitudes as well as their usage behaviors. This suggested that there was something wrong with the tool even as a measure of attitudes.

Fortunately, as soon as we began encountering people who did not seem to match their assigned segment, we started collecting some additional data to help us make sense of the discrepancies. First, we re-screened participants during the visit using the new segmentation tool (the same one they had been screened with initially). We probed to understand both changes in their answers from their original recruiting questionnaire and discrepancies between answers they gave and what we knew about them after spending 3 hours with them. We also screened people using the original segmentation tool, which included slightly fewer attitudinal questions along with a few additional questions on frequency of certain behaviors (of course, its underlying algorithm was different, too, since it had been developed on a different data set.) While the tool algorithm only focused on participants' highest scoring segment, we examined the profile of all their scores across segments. Finally, we experimented with the algorithm to see how changes in respondent answers actually affected their segment assignments.

We considered the possibility that people simply answered the screener in biased ways, guessing at what would get them into the study. Our experience did not support this. Most of the changes in their answers were small, up or down a point on the 7-point rating scales. Changes like these did not seem surprising at all. People who are on the borderline between two scores might lean in different ways at different times just because of chance.

Even large changes in answers can occur strictly because of chance, and not because of any attempt to deceive. When people are asked a survey question that asks them to make a generalization about themselves, and the issue is not one that is a very strong part of their conscious identity, they may answer based on what subset of their experience comes to mind at any given time. (Byer and Holtzblatt [1] discuss the difficulty people have in summarizing their experience as part of their rationale for grounding self-report in observation, in contextual inquiry). Similarly, a recent event that occurred between their initial responses and our re-administration of the tool could skew their responses significantly. Consistent with this, when our participants made large changes in their responses, they were often able to account for them by showing how they had interpreted the question differently at different times or had thought about different sets of their experiences as being relevant to the question at different times.

What was particularly striking was that even the small and unsurprising changes in answers were often enough to change participants' segment assignments. Consistent with this, the small changes in their answers had effects on their segment scores that were relatively large compared to the typical differences between their scores on closely ranked segments. In fact, the changes were often large even compared to the range of scores across their top several segments. (In Figure 2, note how each participant profile has several segment scores that are very close together.)

These observations suggest that the segmentation tool had very low reliability. Reliability is defined as the degree to which the variance in the data is consistently measuring something, whether we know what that "something" is or not. In other words, reliability is a measure of the degree to which the variance in the data is not random noise. That the tool produced

different segment assignments in response to the unsurprising small changes in answers that people gave from one time to the next, essentially randomly, was an indication of low reliability. Paradoxically, many stakeholders for the tool thought that the tool's sensitivity to small changes in answers was an indication of the tool's precision. Apparently, they had never evaluated its test-retest reliability. And of course, the term "precision" only fits if the tool is demonstrably accurate.

The word "accuracy" raises the question of validity. In contrast to reliability, validity has to do with whether we accurately understand what a measurement tool is measuring. The validity of a measurement tool can be no greater than its reliability, because only the non-random part of its variance can be valid. However, high reliability does not ensure high validity.

A consistent bias (as opposed to random variation) introduced into scores represents a validity problem. We found a number of such biases among the segmentation tool items. For example, people who were the least technically sophisticated often thought that any mobile phone is by definition a WiFi device, because it is not connected to a land line. These people were among those most likely to say that having WiFi on their phone was extremely important to them, thus scoring in ways that made them look like more sophisticated users. Similarly, people who used their phones in connection with work tended to answer questions about the importance of the phone in increasing "productivity" in relation to work tasks, which was how the survey authors interpreted the term. However, people who did not use the phone for work often attributed great importance to it in increasing their productivity because it could let them do things like talk to family members while washing dishes, or it could let them take a return call while out shopping, freeing them from having to wait at home for the call. This tended to lead them to be grouped with work users.

What about marketing's claim that the tool was 80% accurate? While stakeholders may have interpreted this as meaning that the tool "got people right," We believe that it simply meant that 80% of people in the original sample were classified the same way by the segmentation tool as they were by the survey as a whole. What had been missing was external validation research, testing it to see if people's scores on the tool predicted something relevant and external to the survey. Our study could have been construed as an attempt at external validation, were it not for the fact that key stakeholders had already bought into the tool's accuracy, so that our study commission took the classification scheme as a given.

As stated above, we also screened people with the older segmentation tool, the one that included a small number of behavioral frequency questions. Most of the attitudinal questions overlapped with those on the new segmentation tool, so we were able to get some indication of where people's original answers would have placed them using the old tool, and where their changed answers placed them. Often, the older tool assigned participants to different segments from the newer tool. In some cases those assignments seemed to better fit our subjective sense of the people. This was interesting, because some at our client's organization believed that the new segmentation tool was better precisely because it left out behavioral

questions. However, as with the new tool, the differences among people's top few segment scores were often tiny, so changes in their answers led to different segment assignments.

As the market research group got more feedback from us about the apparent heterogeneity of their segments, one suggestion they made was to recruit only "core members" of a segment as opposed to "fringe members." In terms of the segmentation tool, these would have been people with a relatively high peak on their rating for that particular segment. However, what if people with true peaks are in fact rare? Figure 2 shows score profiles for 8 participants from segment H. Only one person shows a difference between H and the next highest score that is large compared to his other inter-score differences. Also note the great heterogeneity among profiles, both in their shape and absolute elevation. They have little in common other than that their high point is on H. Our data suggested that core members were either fictional or very rare. Paradoxically, if we had found a few core members, they apparently would have been very unlike most people assigned to the segment.

Lessons Learned

Statistical Groupings versus Individual Measurement

Our experience caused us to think that there is a deeper problem with the idea of developing a segmentation based on massive statistical research and then using a resulting algorithm to label individuals based on the similarity of their response profiles to the average segment profile. Identifying statistical groups and classifying individuals are very different enterprises. However, as sophisticated and

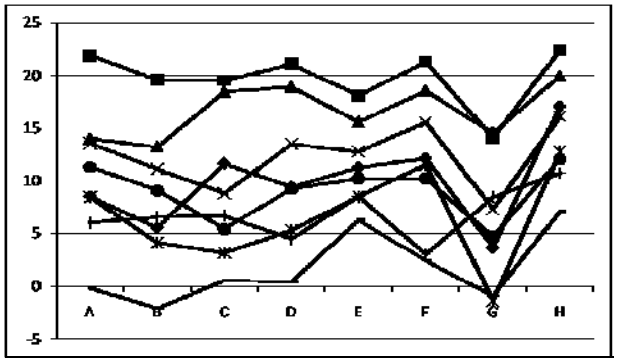

figure 2. Segment score profiles for 8 people assigned to segment H.

"impressive" statistical segmentation and the use of personas to represent segments become increasingly popular in industry, this distinction is being blurred more frequently. This happens when segmentation and/or personas are used as the basis for screening new research participants for small-N studies. As stated earlier, we have seen several examples of this same practice since finishing this project. Also, a number of professional recruiters have told us that they have seen increased use of algorithm-based screeners. They have told us that they "never" find the percentages of people classifying into categories based on these algorithms that clients claim are in the population, that tiny changes in response to the algorithm questions lead to changes in classification, and that people who pass the test of the algorithm frequently disappoint clients because they don't have the characteristics they "should." All of this suggests that our experience is not unique, and that there is an emerging systemic problem in user research practice.

The underlying problem derives from confusing two types of significance. Statistical significance is only about the degree to which the findings were unlikely to occur by chance. In a large enough sample, small average differences can be statistically significant, even when the variance within groups and the overlap among groups is large. Statistical significance is typically enough to make theoretical statements about the influence of particular variables teased out of all the noise. It is also relevant when trying to make incremental improvements by shifting the average value of some measure applied to a large population. But for practical, decision-making purposes in individual cases or in small samples, effect size and likelihood of classification error are more important. In medicine, for example, it might be theoretically interesting to tease out of the noise some evidence that there is a correlation between symptom X and history of behavior Y. But to justify applying an expensive preventative treatment to all individuals who show behavior Y requires a very strong correlation, with low risk of misclassifying individuals. Such strong predictive relationships are rare in human research.

This challenges the prestige of large sample studies in industry. They are likely to detect small differences as statistically significant, when what we often need for robust decision-making in individual cases is big differences. If instead we focus on effect size we will emphasize differences that are big enough to show up convincingly in small samples.

In statistical segmentation research, it is quite possible to find significant differences among groups of people when the groups actually overlap greatly in their key characteristics. In classifying individuals into two statistically different groups, we can be wrong almost 50% of the time, if the original sample was large enough. If we are trying to classify individuals into 8 segments, as was the case here, we will actually be wrong more often than we are right, unless the predictive power is extremely strong. For example, imagine that you are betting on a card drawn from a deck that you know has one extra king. In the long run, you are wisest to bet that your card will be a king, but you will still be wrong most of the time.

Different Uses of Segmentation and Personas

The process of segmentation and persona development naturally accentuates the perception of differences. Whether this is good or bad depends on how things really do cluster in the world, *and* on the purpose of the segmentation. Consider Figures 3 and 4. The open circles represent individuals mapped according to two hypothetical dimensions of difference. The solid circles represent personas. The first diagram represents a world in which things really do fall into fairly neat clusters. The second represents a world with more continuous differences.

If the goal is to inspire design and improve the resonance of marketing messages, then personas work well in both cases. Even in the case of the second diagram, where the personas don't represent a real cluster, they can still do a good job of making the spectrum of diversity concrete in ways that will help guide design. They can stand for regions of the space, without any implication that there is a distinctive species within that space, with clear boundaries around it. In contrast, if the purpose is to develop a model of users that constrains all future research, there is a huge difference between the two diagrams. If the

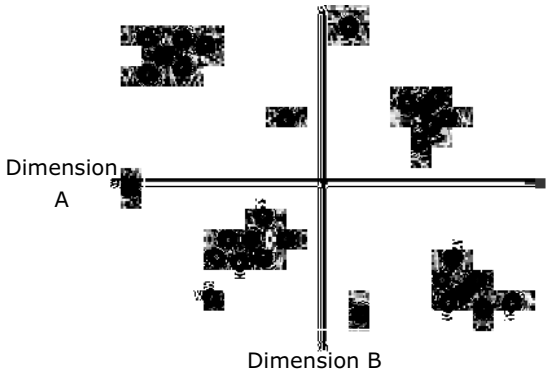

figure 3. If characteristics are discrete

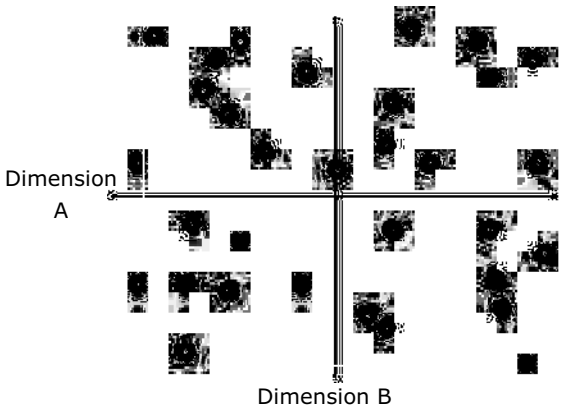

figure 4. If characteristics fall on a continuum

world is more like Figure 4, then trying to recruit people who match the personas in any narrow sense is likely to be a serious mistake. This is why the appearance of precision in algorithm-based recruiting may be deceptive. Trying to find these hypothetical exemplars could be like trying to recruit "average" families with 2.3 children. We should not forget that these are useful abstractions, and do not necessarily reflect distinct species that really exist in nature.

A Better Integration of Quantitative and Qualitative

In a sense, the problem in this case was that the segmentation was prematurely accepted as "truth." If the segmentation based on statistical patterns of survey responses had only been viewed as a hypothesis to be tested with behavioral research, there would have been no problem. What was different in this case was that the stakeholders already "believed" in the segmentation and in the power of the segmentation tool. This is what made it alarming when our data did not support the segmentation. Unfortunately, it may be difficult to remain tentative about the results of a statistical segmentation after you have invested huge amounts of money into it. Nevertheless, iterative cycles of quantitative and qualitative work would progressively lead to more robust segment definitions.

We acknowledge there are several reasons for eventually operationalizing segment definitions in the form of a screening questionnaire. In addition to using such a measurement instrument for screening, it would be needed for market sizing surveys. However, it will always be an empirical question whether clusters defined by behavior and clusters defined by patterns of survey responses map onto each other. To increase the likelihood of this outcome, some of the cycles of research should start with criterion groups known to differ in key behaviors (e.g., smart phone users versus basic phone users, heavy versus light photo users,

mobile web users versus non-users), and then evaluate many possible survey items to see which ones differentiate them the most efficiently. However, we may define much more robust segments using simple indicators rather than highly complex algorithms and extremely nuanced decision rules. While the latter give the appearance of precision they may actually lead to less reliable classification of individuals.

Questioning the Power of Numbers

It is all too easy to allow the mystique of complex statistical research to cloud your thinking and make you forget fundamental issues. Basic principles such as the need to evaluate reliability, and the need for external validation should be respected. Remember that a segmentation based only on similar patterns of survey responses may or may not predict important behavioral distinctions, and this always needs to be evaluated before the segmentation is accepted as "accurate." The power of large sample studies and impressive statistical techniques should never obscure fundamental issues like whether people really behave in ways that are consistent with expressed attitudes and preferences. User experience researchers, even those without advanced statistical training, should be prepared to critique the fundamental logic of the analysis, and whether it really supports the claims and inferences made based on it.

References

[1] Byer, H. and Holtzblatt, K. *Contextual Design: Defining Customer Centered Systems*. Morgan Kaufmann, San Francisco, 1998.

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

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

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

[5] Pruitt, J., and Adlin, T. *The Persona Lifecycle: Keeping People in Mind Throughout Product Design*. Elsevier, New York, USA, 2006.