What Comes to Mind? A Mix of What's Likely and What's Good

Adam Bear[1] (adam.bear@yale.edu), Samantha Bensinger (samantha.bensinger@aya.yale.edu), Julian Jara-Ettinger[1] (julian.jara-ettinger@yale.edu), & Joshua Knobe[2] (joshua.knobe@yale.edu)
[1]Department of Psychology, 2 Hillhouse Avenue, New Haven, CT 06511
[2]Program in Cognitive Science, 2 Hillhouse Avenue, New Haven, CT 06511

Abstract

People can consciously think about only a few things at a time. But what determines the kind of things that come to mind, among a potentially infinite set of possibilities? Two experiments explored whether the things that come to mind are sampled from a probability distribution that combines what people think is statistically likely and what they think is prescriptively good. Experiment 1 found that when people are asked about the first quantities that come to mind for everyday behaviors and events (e.g., hours of TV that a person could watch in a day), they think of values that are proportional to, and intermediate between, what they think is average and what they think is ideal. Experiment 2 quantitatively manipulated distributions of times people devoted to engaging in a novel hobby ("flubbing") and the corresponding distributions of goodness of doing this hobby for various amounts of time. The distribution of values that came to mind resembled the mathematical *product* of the statistical and prescriptive distributions we presented participants, suggesting that something must be both common and good to enter conscious awareness. These results provide insight into the algorithmic process generating people's conscious thoughts and invite new questions about the adaptive value of thinking about things that are both common and good.

Keywords: sampling; consciousness; moral cognition; computation

Introduction

Think of an amount of TV that a person could watch in a day. You might think of one hour, two hours, perhaps even five or six hours. There are no right or wrong answers to this question — it is just a matter of which amount first comes to mind.

This exercise is a contrived example of a computation that the mind performs all the time: selecting samples out of a broad array of possibilities. There are infinitely many possible amounts of TV a person could think about at any given time, but clearly, some amounts are much more likely to come to mind than others. For example, a person will be much more likely to think of three hours than 17.5 hours.

Existing research has explored the ways in which people use sampling algorithms to complete specific aims, such as prediction and decision-making (Stewart, Chater, & Brown, 2006; Vul & Pashler, 2008; Vul, Goodman, Griffiths, & Tenenbaum, 2014). Importantly, however, people also have a capacity to select samples in cases in which they are not explicitly aiming at achieving a specific goal. Even in the absence of a well-defined goal, certain possibilities naturally come to mind while others do not.

In cases of this type, what determines the values that come to mind? One obvious hypothesis would be that the distribution of the amounts that enter people's conscious awareness should mirror people's perception of the corresponding *statistical frequencies* in the population. On this hypothesis, people should have a high probability of thinking of the amounts they perceive as frequent, and a low probability of thinking of the amounts they perceive as infrequent. The mean of the amounts that came to mind would then converge to the amount people regarded as the actual population mean.

But there is also another factor that might play a role here. In the context of decision-making, it is often helpful to be guided by *prescriptive* considerations. In other words, when other things are equal, it is often helpful for people to have a high probability of thinking of options they perceive to be genuinely good, and a low probability of considering options they perceive to be bad.

These two kinds of considerations — statistical and prescriptive — may at first seem to be almost entirely unrelated. However, recent research suggests that there is actually a close connection between the two. People appear to be capable of using a single, undifferentiated representation that mixes together the statistical and the prescriptive (Bear & Knobe, 2017; Icard, Kominsky, & Knobe, 2017; Phillips & Cushman, 2017; Wysocki, 2018). For example, when participants are asked whether a given amount is 'normal,' their answers are influenced both by statistical judgments and by prescriptive judgments. The perceived 'normal amount of TV to watch in a day' is therefore intermediate between the perceived average and the perceived ideal (Bear & Knobe, 2017). Similar results have been obtained across numerous other domains (Wysocki, 2018).

The present paper tests the hypothesis that the distribution of the amounts that enter people's conscious awareness shows this same mix of statistical and prescriptive considerations. On this hypothesis, people are sampling from a distribution that is shaped both by perceived frequency and by perceived goodness.

We consider three basic types of models. First, the probability of thinking of a given possibility might be determined (i) entirely by statistical considerations. Alternatively, it might also be determined by prescriptive considerations. If it is determined by both, it might be that (ii) these two kinds of considerations affect the probability of considering the possibility in a purely additive way (i.e., both considerations independently make things more likely to enter conscious awareness); or (iii) the two kinds of

considerations actually interact in determining the probability of thinking of a possibility. Specifically, the kind of things that most readily come to mind may be things that are *both* statistically likely and prescriptively good, and that neither one of these properties is, on its own, sufficient to bring something to mind.

We explored this question both in a relatively naturalistic setting (Experiment 1) and in an artificial setting in which we could more precisely model people's responses (Experiment 2). We found support for model (iii) from above: what comes to mind is an interaction of what people think is likely and what people think is good.

Experiment 1

In this experiment, we examined how people's intuitions about average and ideal amounts of various ordinary behaviors or events relate to what values spontaneously come to mind. We developed a list of 40 such behaviors or events, 20 of which were borrowed from a similar design from Bear & Knobe (2017). We hypothesized that the values that come to mind would be influenced not only by what was considered average, but also what was considered ideal.

Method

The study proceeded in two parts on Amazon's Mechanical Turk. One set of 100 participants was randomly assigned to judge either the average or ideal value of a set of 20 randomly chosen behaviors or activities, which were randomly selected from the total set of 40. These 20 items were presented in random order to participants. Thus, for 20 of the 40 domains, approximately 50 participants were asked to fill in values like "Average number of hours of TV that a person watches in a day", and approximately 50 other participants were asked to fill in values like "Ideal number of hours of TV for a person to watch in a day". To avoid demand characteristics, participants were only always asked about either averages or ideals, never both in the same session.

A separate group of 100 subjects participated in the sampling part of the experiment, in which they gave values that first came to mind. Participants were instructed to simply "enter the first number that comes to mind" when reading the presented phrase, and it was emphasized that there was no "correct" answer. In order to encourage participants to give a spontaneous judgment, we instructed them to try to give each response in under 5 seconds. However, responses were still solicited after this time delay. After completing two practice trials, the participants were presented with a random 20 out of 40 domains, presented in random order. Each page simply displayed a phrase like "NUMBER OF HOURS OF TV FOR A PERSON TO WATCH IN A DAY" and a timer counting down from 5 seconds, along with a box for subjects to give their response.

Results

Participants' responses in each condition were averaged for each of our 40 domains (Table 1). Responses from participants who failed an attention check or that were 3 standard deviations away from the mean answer for a given question were excluded.

Since our questions asked about very different kinds of quantities (hours, calories, etc.), assumptions of normality were violated. To address this problem, mean responses for each measure were converted to log scale.

To examine how judgments of averages and ideals affect sampling judgments, we compared a regression model in which only average judgments predict sampling judgments to a model in which both average and ideal judgments predict these judgments. The latter model reveals that both judged averages, $\beta = .77$, $SE = .05$, $p < .001$, and judged ideals, $\beta = .18$, $SE = .04$, $p < .001$, significantly predict sampling judgments. Moreover, the corrected Akaike Information Criterion (AICc) for this model (17.75) is markedly lower than that for a model in which only judged averages predict normality judgments (30.88), suggesting that it is a more appropriate model of the observed data. Following Wagenmakers & Farrell (2004), the strength of evidence in favor of the more complex model can be quantified with an evidence ratio. This ratio was 709, indicating a highly favorable fit for the model that includes ideal judgments as a predictor.

We also conducted non-parametric analyses to explore whether people's samples were intermediate between judged averages and ideals. For a given sample to be intermediate, it must be both on the ideal side of the average and the average side of the ideal. For the 40 domains, 29 were on the ideal side of average (binomial $p = .006$), and 37 were on the average side of ideal (binomial $p < .001$). Further, 26 out of 40 of the sample values met both of these criteria — i.e., they were intermediate between average and ideal judgments. Thus, although many sample values were not intermediate, the proportion that were intermediate was considerably greater than what would be expected by chance (binomial $p < .001$ with a null hypothesis of 1/3, since there are two possible ways that an item can be non-intermediate).

Discussion

In this experiment, the values that spontaneously came to people's minds, like hours of television watching, were best explained by considering both statistical reasoning (what is considered average) and prescriptive judgments (what is considered ideal). However, this result does not tell us about the computational process that generated these judgments. In the next experiment, we explore this question in more detail.

Experiment 2

Experiment 1 found that what comes to mind depends on both statistical and prescriptive kinds of information. But because people's statistical and prescriptive beliefs were

collapsed into single judgments of average and ideal, respectively, we could not get a detailed understanding of how this information was being used to produce people's sample judgments.

In Experiment 2, we moved away from simple point values of "average" and "ideal" to a more controlled setting, in which the entire distributions of statistical and prescriptive information that participants were exposed to were varied, so we could explore how these full distributions were functionally combined to produce samples that came to mind. In particular, we could compare models in which the samples were a function of a weighted *sum* of statistical frequency and prescriptive goodness to models in which the samples were a function of the *product* of these two types of information.

Method

Four-hundred participants from Amazon's Mechanical Turk were randomly assigned into one of four conditions in a 2 x 2 design. We orthogonally manipulated the statistical distribution of values presented to participants (unimodal vs. bimodal) and whether our fictional hobby was healthy or unhealthy (high ideal vs. low ideal).

Participants were first presented with a description of the fictional hobby of "flubbing". In the low ideal condition, participants were told that "although it is safe to flub for a few minutes every week, doctors warn that there are serious health risks associated with flubbing for longer periods of time." The high ideal condition, in contrast, stated that "doctors advise their patients to flub as much as possible" and that the more people flub, the healthier they are.

Participants were then told that they would be presented with amounts of time (in minutes) that 100 people flubbed in the past week (one at a time, on separate pages), along with health grades, ranging from A+ to D-, that indicated the healthiness of flubbing for each of these amounts of time.

Grades were calculated in the following way. In the high ideal condition, all amounts of flubbing greater than 80 minutes were given an A+, and all amounts less than 20 were given a D-. The opposite was the case in the low ideal condition. Then, within the 20–80 range, grades were spaced linearly in intervals of 5, such that 75–80 corresponded to A+, 70–75 A-, and so on for the high ideal condition, and the reverse for the low ideal condition.

The amounts of flubbing were sampled from a normal distribution with μ = 45 and σ = 15 in the unimodal condition and a sum of normal distributions with μ = 35 and 75, and σ = 5, in the bimodal condition. These values were rounded to the nearest integer. Within each of these conditions, all participants were given the exact same 100 values (i.e., we only sampled from these distributions once per condition), presented in a different random order for each participant.

After viewing all 100 values of flubbing, participants were asked, without forewarning, what was the first number of minutes of flubbing that came to mind. As in Experiment 1, they were told that there was no need to deliberate about this and that we were not looking for a particular correct answer. Participants were also asked afterwards what they thought the average amount was.

Computational Framework

To investigate how prescriptive information influences participants' sample judgments, we consider several models that combine statistical and prescriptive information to produce a probability distribution of possible samples. On a simple account, people might simply draw samples from the statistical distribution that generated the data, ignoring information about goodness. Thus, participants' samples may be guided by

$$\text{Control}(x;C) = \text{Stat}(x) + C, \quad (1)$$

where C is constant term to account for any uniform baseline probability of sampling. In contrast, there are several ways in which prescriptive information could combine with statistical information to play a role in predicting what comes to mind. (To save space, we ignore the obviously false model in which *only* prescriptive considerations influence samples.)

We focus on two sets of possible models. First, what comes to mind may be a weighted combination, or sum, of statistical and prescriptive information, such that amounts that are more common are more likely to be sampled, and amounts that are more desirable are also more likely to be sampled, but the interaction between these two pieces of information does not play any role. In other words, statistical frequency and normative goodness may simply be two independent factors that contribute to a value's probability of coming to mind. If so, then participants' judgments should follow a distribution of the form

$$\text{Add}(x;c,b,w_1,w_2,C) = w_1*\text{Stat}(x) + w_2*\text{Ideal}(x;c,b) + C, \quad (2)$$

where w_1 and w_2 are weighting parameters.

Alternatively, values may only (or primarily) come to mind when they are both statistically frequent *and* normatively good. That is, the probability of sampling may be proportional to the *product* of frequency and goodness. If so, participant judgments should follow a distribution fit by

$$\text{Mult}(x;c,b,C) = \text{Stat}(x)*\text{Ideal}(x;c,b) + C. \quad (3)$$

In the additive and multiplicative models (Eqs 2 and 3, respectively), we consider two potentially relevant factors in mapping the letter grades we presented participants to the Ideal function: baseline goodness (parameter b) and convexity (parameter c). At baseline, bad amounts of flubbing (e.g., 10 minutes in the high ideal condition) might be given an ideal value of 0 or might be given a value that is only slightly worse than good amounts of flubbing (e.g., 90 minutes in the high ideal condition). The baseline goodness

parameter, therefore, tracks how bad people think it is to flub an amount that deviates greatly from the highest possible grade.

The convexity of the function tracks the relative drop-off in goodness as values move away from the highest possible grade. For example, people may think that a B grade is much worse than an A grade, but a D grade is only slightly worse than a C grade, indicating a highly convex function. In contrast, convexity of 0 would correspond to a completely linear function, in which the difference in goodness between A and B is the same as that between C and D. (We ignore the possibility of concavity here, as our data strongly suggests that goodness takes a convex form.)

Putting this all together and assuming that goodness cannot get higher than 1, we model goodness in the high ideal condition as

$$\text{Ideal}(x;c,b) = e^{c(x-80)} + b \qquad (4)$$

when $x < 80$ and 1 otherwise. The low ideal condition simply flips the function, such that $x - 80$ becomes $20 - x$ for values of $x > 20$, with $\text{Ideal}(x) = 1$ when $x < 20$.

Results

We first explored whether participants' sample judgments of what comes to mind were influenced by the goodness of the hobby (i.e., whether it was healthy or unhealthy). This was confirmed: participants' samples averaged 33.78 across low ideal conditions and 60.47 across high ideal conditions, $t(394) = 15.13$, $p < .001$. (Note that a few observations were lost by participants who did not give numerical responses to our sample question, explaining the 4 missing observations.) Moreover, these judgments deviated significantly more in the direction of the ideal than participants' estimations of the average amount of flubbing they saw: judgments of average were 40.77 and 49.48 in the low and high ideal conditions, respectively ($t(198) = 5.93$, $p < .001$ and $t(195) = 4.83$, $p < .001$ comparing average and sample judgments in each condition).

Next, we fit each class of model to our observed data using maximum likelihood estimation, implemented with MATLAB's *fmincon* function. All parameters were constrained to be between 0 and 1, and all models fit to the entire dataset (i.e., one set of parameters were fit to all conditions together).

Despite the additive model's two extra free parameters, the multiplicative model outperformed it: the best-fitting multiplicative model had a negative log likelihood (NLL) of 1,567, while the best-fitting additive model had a NLL of 1,642. Moreover, a comparison of AIC values suggests that the multiplicative model is definitively more likely to minimize information loss (evidence ratio $> 10^{32}$). As expected, the control model performed much worse than both of these models, with a NLL of 1,674, and had a much worse AIC than the best-fitting multiplicative model (evidence ratio in favor of multiplicative model $> 10^{44}$).

Figure 1 presents histograms of the sample values across the four conditions, along with model fits (black lines) from the best-fitting multiplicative model, with $c = 0.084$, $b = 0.047$, and $C = 0.0005$. The green lines indicate the ideal function for each of these conditions with the fitted parameters, and the gray lines displays the statistical distributions. It is clear from visual inspection that the multiplicative model does a fairly good job of characterizing participants' sample judgments, particularly in the bimodal conditions. In the unimodal conditions, it spreads out the probability mass across less ideal values slightly more than observed, but still captures the general shape.

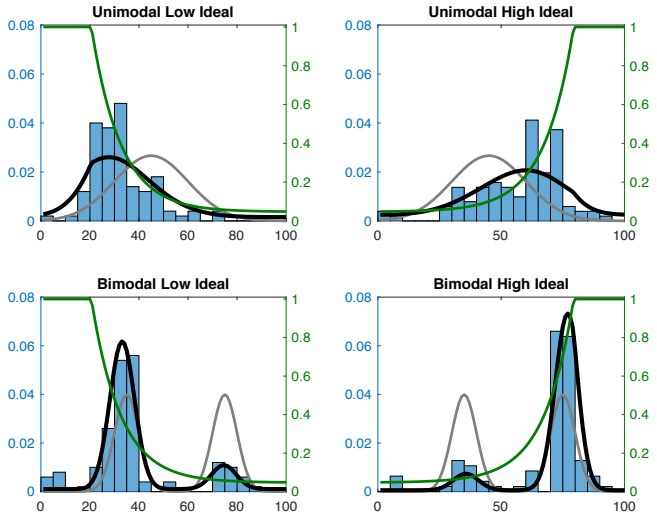

Figure 1: Data and model fits from Experiment 2. Vertical bars show proportion of values sampled by participants. Gray and green lines show the statistical information and prescriptive information (fit through Eq. 4) that participants saw, and the black lines show the predictions of the best multiplicative model (Eq. 3).

Control Experiment

Experiment 2 confirmed that the normative information presented to participants strongly influenced what came to mind. However, it is possible that *any* information that we would give to participants other than statistical frequency could have a similar effect. If so, our data would not provide evidence that goodness, in particular, exerts an influence on sample judgments, but just any extraneous information.

To address this worry, we conducted a ($N = 101$) follow-up experiment on Mechanical Turk, based on the distribution and ideal values from the unimodal, low ideal condition above. In the experimental condition, participants were told (as above) that too much flubbing was bad for their health. However, instead of presenting health grades, we presented descriptions of "Very Good," "Somewhat Good," "Somewhat Bad," and "Very Bad" to indicate the healthiness of the flubbing amounts, which corresponded to grades in A-range, B-range, C-range, and D-range, respectively.

In the control condition, we instead told participants that flubbing is a hobby that people "like to do at various altitudes" and that people tend to flub for longer amounts of time at lower altitudes, but the altitudes do not influence how enjoyable flubbing is. Then, using the same values from the other condition, we presented participants with descriptions of the altitudes at which different people flubbed ("Very High," "Somewhat High," "Somewhat Low," and "Very Low"), which corresponded exactly to the mapping of labels from the other condition.

Our results confirmed that the prescriptive information from the experimental condition exerted a larger influence on people's sample judgments (μ = 34.59) than the irrelevant altitude information (μ = 41.16), $t(99) = 2.11$, $p = .038$, suggesting that prescriptive information plays a specific role in biasing what comes to mind. Moreover, participants' sample judgments in the altitude condition were not significantly different from their judgments of the average (μ = 45.06), $t(49) = 1.43$, $p = .160$.

Discussion

In this experiment, we manipulated the distributions of statistical and prescriptive information we presented to participants to more quantitatively measure the extent to which this information influences what comes to people's minds. Consistent with Experiment 1, we found that goodness exerts a strong influence on people's sample judgments — and this is not true for other irrelevant information. More importantly, we found that people sample proportional to the *product* of statistical frequency and prescriptive goodness, suggesting that, in general, something needs to be both common and desirable for it to come to mind.

General Discussion

Two experiments found that the possibilities that spontaneously enter people's minds are a mixture of what is thought to be likely and what is thought to be ideal. These results provide direct support for past work that has indirectly suggested such an influence of prescriptive considerations on the possibilities that people consider (Bear & Knobe, 2017; Icard et al., 2017; Phillips & Cushman, 2017; Wysocki, 2018). Moreover, Experiment 2 provided novel support for a computational account of how the mind combines statistical and prescriptive considerations to produce the quantities that enter conscious awareness: frequency and goodness seem to be *multiplied*.

A key question for further research will be why people sample possibilities from a distribution that is shaped in this way by both statistical and prescriptive considerations. In answering this question, one strategy would be to posit one reason why it would be adaptive for people to think about possibilities that arise frequently and then another, completely unrelated reason why it would be adaptive for people to think about possibilities they regard as good. For example, perhaps it is adaptive for people to think about actions that are performed frequently because people are likely to find themselves in situations in which someone else is performing one of those actions. Then, unrelatedly, perhaps it is adaptive for people to think about actions they regard as good because they will need to consider those actions in their own planning or deliberation. This type of hypothesis is certainly a plausible one, but it faces at least some difficulty in explaining why people specifically tend to think about possibilities that are *both* frequent and good.

A second strategy would be to develop a more unified account that explains why it might be adaptive to think about possibilities in this hybrid way. For example, of all of the infinitely many things you might choose to do in your life, which would be worthy of further conscious consideration this afternoon? Given that you could not feasibly consider *all* candidate actions, a useful heuristic might be to consider, from among the reasonably common actions, those that seem relatively good. Thus, one would not consider options that are extremely infrequent (e.g., traveling to the moon) or options that seem extremely bad (e.g., robbing a bank) but only options that seem to be both reasonably frequent and also relatively good. On this hypothesis, statistical and prescriptive considerations are relevant for the same reason, namely, that they help in identifying options that might be worth considering in deliberation.

In sum, the present work offers a first step in describing the informational and computational factors that contribute to a largely unexplored psychological phenomenon: thoughts entering conscious awareness. The results suggest that this process involves a surprisingly systematic blend of statistical and prescriptive considerations. Further work should continue to explore the nature of this blending and its role in downstream cognitive processes.

References

Bear, A. & Knobe, J. (2017). Normality: Part descriptive, part prescriptive. *Cognition, 167*, 25–37.

Gershman, S. J., Vul, E., & Tenenbaum, J. B. (2012). Multistability and perceptual inference. *Neural Computation*, *24*, 1–24.

Icard, T. F., Kominsky, J. F., & Knobe, J. (2017). Normality and actual causal strength. *Cognition*, *161*, 80–93.

Morris, A., Phillips, J., & Cushman, F. (2018). *Value-guided choice sets support efficient planning*. Manuscript in preparation.

Phillips, J. & Cushman, F. (2017). Morality constrains the default representation of what is possible. *Proceedings of the National Academy of Sciences*, *114*, 4649–4654.

Stewart, N., Chater, N., & Brown, G. D. (2006). Decision by sampling. *Cognitive Psychology*, *53*, 1–26.

Vul, E. & Pashler, H. (2008). Measuring the crowd within: Probabilistic representations within individuals. *Psychological Science*, *19*, 645–647.

Vul, E., Goodman, N., Griffiths, T. L., & Tenenbaum, J. B. (2014). One and done? Optimal decisions from very few samples. *Cognitive Science*, *38*, 599–637.

Wagenmakers, E. J. & Farrell, S. (2004). AIC model selection using Akaike weights. *Psychonomic Bulletin & Review*, *11*, 192–196.
Wysocki, T. (2018). *Normality: A two-faced concept.* Unpublished manuscript.

Table 1: Mean Average (A), Ideal (I), and Sample (S) Judgments across Domains from Experiment 1

Domain	**A**	**I**	**S**	**Domain**	**A**	**I**	**S**
Hours TV/day	3.38	1.63	2.87	*Drinks frat bro consumes/wkend*	11.12	6.63	15.64
Sugary drinks/wk	9.17	2.41	5.91	*Times honk at drivers/wk*	2.67	0.72	2.53
Hours Exercise/wk	4.00	5.58	6.33	*Mins on social media/day*	60.57	35.40	59.10
Cals consumed/day	2225.91	1900.00	1859.24	*Times parent punishes child/month*	6.58	2.28	3.25
Servings fruits & veggies/month	40.00	94.96	39.16	*Miles walked/wk*	9.79	12.96	9.96
Lies told/wk	9.57	1.17	8.44	*% people drive drunk*	11.30	1.23	9.45
Mins late for appointment	14.22	3.04	13.60	*Times cheat on partner in life*	1.52	0.00	1.73
Books read/yr	7.22	17.40	8.45	*Times snooze alarm/day*	2.13	0.76	1.98
Romantic partners in life	6.09	5.77	8.06	*Parking tickets/yr*	1.67	0.04	1.37
Country's international conflicts/decade	11.67	1.36	4.15	*Times car wash/yr*	10.77	12.85	11.31
$ cheated on taxes	437.45	82.00	350.32	*Cups coffee/day*	2.21	1.84	2.72
% students cheat on HS exam	33.00	2.17	19.50	*Desserts/wk*	3.85	2.92	4.04
Times checking phone/day	28.57	7.68	16.57	*Loads of laundry/wk*	3.42	2.70	3.75
Mins waiting on phone for customer service	20.21	3.88	13.29	*% smokers*	22.81	6.16	20.79
Times called parents/month	5.00	5.50	7.04	*% HS students underage drink*	35.81	13.71	32.96
Times clean home/month	5.78	4.35	6.24	*% lie on dating website*	50.56	13.40	47.20
Times computer crash/wk	3.07	0.12	1.14	*Servings carbs/day*	62.43	16.13	33.23
% HS dropouts	10.67	1.29	11.49	*Txt msgs sent/day*	27.18	12.88	18.10
% middle schoolers bullied	17.59	0.81	19.46	*Times lose temper/wk*	2.60	0.56	2.20
Hrs slept/night	6.69	7.84	7.32	*Times swearing/day*	8.69	5.88	11.26