Research Report

Psychological Science
24(2) 219–224

DOI: 10.1177/0956797612459762
http://pss.sagepub.com

Daily Horizons: Evidence of Narrow Bracketing in Judgment From 10 Years of M.B.A. Admissions Interviews

Uri Simonsohn[1] and Francesca Gino[2]
[1]The Wharton School, University of Pennsylvania, and [2]Harvard Business School, Harvard University

Abstract

Many professionals, from auditors, venture capitalists, and lawyers, to clinical psychologists and journal editors, divide continuous flows of judgments into subsets. College admissions interviewers, for instance, evaluate but a handful of applicants a day. We conjectured that in such situations, individuals engage in narrow bracketing, assessing each subset in isolation and then—for any given subset—avoiding much deviation from the expected overall distribution of judgments. For instance, an interviewer who has already highly recommended three applicants on a given day may be reluctant to do the same for a fourth applicant. Data from more than 9,000 M.B.A. interviews supported this prediction. Auxiliary analyses suggest that contrast effects and nonrandom scheduling of interviews are unlikely alternative explanations of the observed pattern of results.

Keywords

decision making, judgment, heuristics

Received 10/12/11; Revision accepted 5/20/12

Many professionals, from auditors, venture capitalists, and lawyers, to clinical psychologists and journal editors, have jobs that involve a continuous flow of judgments that they execute, over time, in small subsets. University admissions officers, for example, interview hundreds of applicants per year in subsets of a handful each day. These arbitrarily created subsets should have no influence on experts' judgments. Although the merit of an M.B.A. applicant may partially depend on the pool of applicants in that year, it should not depend on the few others who happen to be interviewed the same day. However, decision makers often engage in *narrow bracketing*—that is, they fail to integrate the consequences of many similar decisions (for a review, see Read, Loewenstein, & Rabin, 1999).

In this article, we examine how narrow bracketing affects judgments. Research has shown that people focus too much on the particular case at hand and neglect background information (Brenner, Griffin, & Koehler, 2005; Griffin & Tversky, 1992; Massey & Wu, 2005), and we hypothesized, along similar lines, that when people conduct a subset of judgments, they do not sufficiently consider the other subsets they have already made or will make in the future. Considering that people exaggerate the extent to which small samples resemble large samples (Tversky & Kahneman, 1971), we reasoned that people avoid making subsets of judgments that deviate much from what they expect the overall set of judgments to be like. For instance, an interviewer who expects to evaluate about 50% of applicants in a pool positively may be reluctant to evaluate much more or less than 50% of applicants positively on any given day. An applicant who happens to interview on a day when several others have already received a positive evaluation would, therefore, be at a disadvantage.

We tested this prediction by analyzing 10 years of data on M.B.A. applications to an American business school (with which neither of us is affiliated), assessing whether applicants' scores were negatively correlated with the average score of previous applicants who were interviewed on the same day. We studied narrow bracketing in the context of experts working in their everyday environment, rather than in the laboratory, in order to avoid the possibility that judgments would be negatively autocorrelated as a result of changes in beliefs about the distribution of underlying quality (e.g., "I rated the previous three applicants very highly, so the pool must be strong; I will start evaluating more harshly") or in scale use (e.g., "I am giving too many high scores, so I must be rating

Corresponding Author:
Uri Simonsohn, University of Pennsylvania, The Wharton School, 3730 Walnut St., 500 Huntsman Hall, Philadelphia, PA 19104
E-mail: uws@wharton.upenn.edu

too leniently; I will start evaluating more harshly"). Because experts who interview large numbers of applicants year after year should not revise their beliefs about the quality of the applicant pool or change their scale usage upon seeing a handful of weak or strong applicants on a single day, they are an ideal testing ground for studying the consequences of narrow bracketing in judgment.

Empirical Analyses

Data

Our data set consisted of 14,065 interviews of M.B.A. applicants between 2000 and 2009.[1] After erroneous or incomplete entries and interviews conducted by alumni were eliminated, the sample contained 9,323 interviews conducted by 31 interviewers. Following the disclosure guidelines recommended by Simmons, Nelson, and Simonsohn (2011), we have provided the full list of variables (along with information on data cleaning) in the Supplemental Material available online. Given the nested nature of the data, all analyses clustered standard errors at the interviewer level (i.e., they took into account that we had repeated measures for interviewers).

Interviewers rated applicants (scales from 1 to 5) on (a) communication skills, (b) being driven, (c) ability to work in teams, (d) being accomplished, and (e) interest in the school. They also provided an overall evaluation (scale from 1 to 5) of each interview. We refer to the latter as *scores* and the former as *subscores*.

Interviewers conducted an average of 4.5 interviews per day (SD = 1.9) on those days that they conducted interviews, and gave an average score of 2.8 (SD = 0.9). The data set included information about both the applicants (e.g., their GMAT scores) and the interviews (e.g., date and time).

Main results

We estimated regressions with applicant's interview score as the dependent variable and the average score given to previous applicants by the same interviewer earlier on the same day as the key predictor. We controlled for characteristics of the applicant and interview and for fixed effects of the interviewer. Analyses were restricted to the third and later interviews for a given interviewer on a given day. Results are presented in Table 1.

Our baseline model (Model 1 in Table 1) controlled only for interviewer effects (with 30 binary variables that allowed estimation of a separate main effect for each of the 31 interviewers) and for month and year of the interview (allowing a main effect for each month of each year in the sample). The point estimate for the impact of the average score of previous interviews was, as predicted, negative and significant, b = −0.116, p = .005.

Model 2 added controls for an applicant's characteristics, and Model 3 added controls for an interview's characteristics. The point estimate of interest was still negative and significant in both models. In Model 4, we added the score given to an applicant's written application, which would be expected to control for many other unobservable differences across applicants. The point estimate of interest remained negative and significant, b = −0.088, p = .018.[2] The stability of the key point estimates when we added controls into the regression gives us confidence that the main finding was not the result of omitted variables.[3]

Figure 1 depicts the residuals from a regression that controlled for all observable variables in Table 1 except the key predictor of interest, the average score given to previous applicants on the same day. The graph suggests that modeling the effect as linear and symmetric for high and low average scores is reasonable.[4]

Effect size. The key point estimates in Table 1 ($b \approx -0.1$) imply that as the average score of previous applicants on a given day (SD = 0.75) increased by 1 standard deviation, the expected score for the next applicant dropped by about 0.075. To counteract such a decrease, an applicant would need 30 more points on the GMAT, 23 more months of experience, or 0.23 more points in the score for the written application.

Another benchmark for the effect size comes from the interview subscores. We conducted a regression that predicted the overall score from these five subscores; all covariates from Table 1 except average previous score were included. Results indicated that the effect size of 0.075 was equivalent to the interviewee increasing his or her communication rating in the interview by about 0.33 standard deviations, or increasing his or her subscore for interest in the school by 0.89 standard deviations.[5]

Heterogeneity. We also considered whether the effect of interest exhibited heterogeneity across interviewers and heterogeneity within interviewers across days. To evaluate heterogeneity across interviewers, we estimated the full specification (Model 4 in Table 1) for each interviewer separately. Of the 31 interviewers in the sample, 18 had enough interviews to allow such estimation given the large number of predictors. For all but 1 of these 18 interviewers, the point estimate for the effect of the average score given previously on the same day was negative. Eight of these 17 negative estimates were significant at the .05 level. The single positive point estimate was not significant (p = .496). The aggregate pattern we observed, then, was not driven by a small subset of interviewers.

We studied heterogeneity within interviewer across days following recommendations by two anonymous referees who suggested that variability in previous scores during a day may influence the impact of the average score. For example, an interviewer might be more reluctant to give an interviewee a 4 after rating three candidates in a row with a 4 than after rating three candidates with a 3, 4, and 5, respectively. Consistent

Table 1. Point Estimates From Regressions Predicting the Current Interviewee's Score and Placebo Variables

	Predicting the current interviewee's score				Placebo tests	
Predictor	Model 1 (baseline)	Model 2 (interviewee controls)	Model 3 (interviewee and interview controls)	Model 4 (interviewee and interview controls; written-application score added)	Model 5 (predicting GMAT score)	Model 6 (predicting experience in months)
Average score given by the same interviewer to previous applicants that day	−0.116**	−0.110**	−0.105**	−0.088*	0.089	0.250
	(0.038)	(0.035)	(0.036)	(0.035)	(2.062)	(0.958)
GMAT score of applicant[a]		0.244**	0.250**	0.079*	—	1.140*
		(0.036)	(0.035)	(0.032)		(0.495)
Job experience of applicant (months)[a]		0.324**	0.319**	0.254**	10.356*	—
		(0.057)	(0.055)	(0.055)	(4.540)	
Number of interviews by same interviewer that day						
Total			−0.000	0.001	0.844	0.504†
			(0.012)	(0.012)	(0.677)	(0.291)
Before the current interview			−0.018	−0.010	−0.460	0.008
			(0.013)	(0.014)	(1.275)	(0.360)
Score given by reader of application				0.340**	24.193**	2.101**
				(0.044)	(1.719)	(0.492)
Number of observations	4,456	4,312	4,312	3,754	3,754	3,754
R^2	.322	.381	.387	.484	.288	.726

Note: The sample consisted of interviews of M.B.A. applicants from 2000 through 2009; the interviews were given by 31 interviewers working for the admissions office of a private business school. Analyses were restricted to the third and later interviews for a given interviewer on a given day. The table presents point estimates from ordinary least squares regressions. Standard errors, clustered by interviewer, are given in parentheses. All models included month × year dummies (k = 12 × 9) and interviewer dummies (k = 31), where k indexes the degrees of freedom when k is greater than 1. *Model 2* controlled for interviewee's gender, race (k = 9), age, and age-squared. *Models 3* and *4* also controlled for the hour of the interview (k = 12) and the interview's location (k = 4). Models 5 and 6 have the same covariates as Model 4.

[a]GMAT score and job experience were divided by 100 to arrive at readable point estimates.

†p < .10. *p < .05. **p < .01.

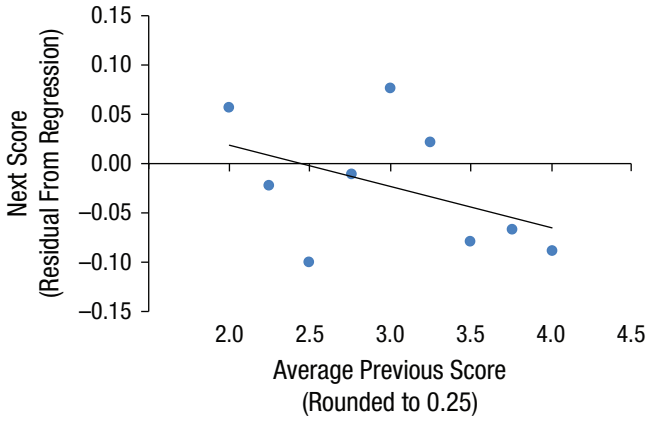

Fig. 1. Relationship between the average previous score given by an interviewer on a given day and the next score given by that interviewer. The data plotted are residuals from Model 4 (see Table 1), an ordinary least squares regression that included applicant's characteristics, interview's characteristics, and applicant's score on the written application as covariates.

with this prediction, analyses revealed that the effect of previous scores was twice as large following a set of identical scores as following a set of heterogeneous ones ($b = -0.111$ vs. $b = -0.059$). Despite the dramatic difference in these point estimates, it was not statistically significant; we lacked the power to detect sensibly sized effects.

Possible alternative mechanisms

We considered two alternative explanations for our main findings. First, there might have been a contrast effect. If interviewers employed recently seen applicants as a reference, then applicants following stronger applicants would have seemed weaker to interviewers, and applicants following weaker ones would have seemed stronger. Such an effect would have led to a negative correlation among ratings within a day. Second, nonrandom sequencing of applicants could also have led to our findings: If stronger candidates tended to be followed by weaker ones in the daily scheduling of interviews (or vice versa), this also would have resulted in a negative correlation among ratings within a day. We address each of these explanations next.

Was there a contrast effect? We tested two sets of predictions that allowed us to evaluate whether our findings were due to a contrast effect, rather than narrow bracketing. The first set of analyses involved interview subscores. Recall that each applicant was rated on five specific attributes (e.g., communication skills) in addition to receiving the holistic overall score. Because the evaluation of these specific attributes was more perception based and specific, one might expect the subscores to be more susceptible to contrast effects than the overall score was; if our core finding was driven by a contrast effect, these subscores should also show an (arguably more pronounced) effect. For example, the contrast between an eloquent applicant and an inarticulate one seen back to back should have been starker than the contrast between applicants who differed in their overall strength aggregated across a broad range of attributes. Moreover, research on contrast effects in person perception shows that such contrasts occur only through specific and relevant attributes (Higgins, Rholes, & Jones, 1977; Srull & Wyer, 1979), which suggests that an overall contrast effect is likely to be the downstream consequence of specific attribute contrasts, and hence that the latter are a necessary condition for the former.

The opposite prediction follows from the narrow-bracketing account. Because interviewers are unlikely to be concerned about keeping a balanced distribution of each subscore, and they may even have difficulty remembering the subscores they gave to previous applicants, this account indicates that subscores should be influenced weakly, if at all, by previous subscores.

The second set of analyses involved the moderating role of how far into the set of daily evaluations an interviewer was. If interviewers engaged in narrow bracketing, then as a day was about to end, imbalances should have been particularly aversive, and the inclination to respond to the previous ratings should have been stronger. The contrast-effects literature on person perception is too nuanced to make an unambiguous prediction regarding the effect of an interview's serial position within a day. For example, depending on whether interviewers were focusing on similarity or differences among candidates, or whether previous candidates were at extreme or moderate levels on a given attribute, one would expect contrast effects to get stronger or weaker, or even to become assimilation effects, as the day progressed (for a review, see Wheeler & Petty, 2001)

This second set of analyses is hence asymmetric (Larrick & Wu, 2007). If the impact of previous ratings did not increase as the day was about to end, the narrow-bracketing account would be an inadequate explanation for the data. However, if the effect of previous scores did get stronger toward the end of the day, a contrast-effect mechanism would not be ruled out.

We conducted regressions analogous to those reported in Table 1 on the subscores. For example, we estimated the impact of the average communication-skill score given to previous applicants on a given day by a given interviewer on the communication score given to the current applicant. For all five subscores, the effect was weak and not significant—communication skills: $b = -0.009$, $p = .748$; drive: $b = -0.037$, $p = .279$; ability to work in teams: $b = -0.027$, $p = .411$; accomplishment: $b = -0.052$, $p = .073$; and interest in the school: $b = 0.025$, $p = .531$. To reduce noise, we also averaged the five subscores and conducted the analyses on that average as if it were a sixth subscore, again obtaining an insignificant effect, $b = -0.051$, $p = .223$. Because subscores do not show the same effect the overall score does, we conclude that contrast effects are an unlikely explanation for our findings.

As noted by an anonymous referee, the greater specificity of the subscores may make them insufficiently ambiguous to exhibit biases. Providing a definitive answer to this concern

would require unavailable data on the relative ambiguity of subscores versus overall scores in the minds of the interviewers. The subscores in the data set, however, seem to us to be about as ambiguous as the ratings used by scholars examining priming effects in person perception (e.g., assessing if Donald is kind or reckless; see Thompson, Roman, Moskowitz, Chaiken, & Bargh, 1994; Winter & Uleman, 1984).

To assess the moderating role of approaching the end of the day, we estimated regressions separately for subsets of interviews occurring in particular serial positions within the day. The point estimates of interest, those for the effect of the average previous score, are plotted in Figure 2. As predicted, the impact of previous scores grew larger and became significant as a day progressed.

Was the objective strength of applicants negatively serially correlated within day? We estimated regressions predicting applicants' GMAT scores and job experience from the average scores given to previous interviewees on the same day. These were, in effect, placebo tests: If our interpretation of the data is correct, the average previous interview score on the same day would not be expected to predict these other dependent variables. Models 5 and 6 of Table 1 show small, positive, and nonsignificant effects of previous scores in these placebo tests. This evidence is hence inconsistent with candidates' objective strength accounting for our core finding.

General Discussion

Building on the choice-bracketing literature, which shows that decision makers insufficiently take into account the aggregate consequences of many similar decisions (Read et al., 1999), we examined narrow bracketing in judgment. In line with research showing that individuals put too much weight on an individual case and too little on background information (Brenner et al., 2005; Griffin & Tversky, 1992; Massey & Wu, 2005), we conjectured that people conducting sequences of subsets of judgments insufficiently take into account other judgments they have made prior to the current subset or will make in the future. As a result, people avoid generating subsets of judgments that deviate much from the expected overall distribution. We found support for this prediction using data from more than 9,000 interviews of M.B.A. applicants. Our analyses suggest that the evidence is inconsistent with nonrandom scheduling of interviews or sequential contrast effects.

Although we have focused on well-defined daily subsets, a similar bias may occur when people conduct larger sets of evaluations and generate subsets spontaneously in their minds. Imagine, for example, a judge who must make dozens of judgments a day. Given that people underestimate the presence of streaks in random sequences (Gilovich, Vallone, & Tversky, 1985), the judge may be disproportionately reluctant to evaluate four, five, or six people in a row in too similar a fashion, even though that "subset" was formed post hoc.

We propose three specific mechanisms by which narrow bracketing in judgment may account for our findings. The first is based on the belief in the law of small numbers (Tversky & Kahneman, 1971): Upon giving a set of positive judgments or a set of negative judgments, interviewers may form an expectation that a weaker or stronger candidate, respectively, "is due." Or they may attempt to correct for perceived errors in their ratings based on deviations from expectations regarding the distribution of ratings. The second possibility is that interviewers engage in mental accounting (Thaler, 1985, 1999), simplifying the task of maintaining a given long-term target of positive evaluations by applying their target to each "daily

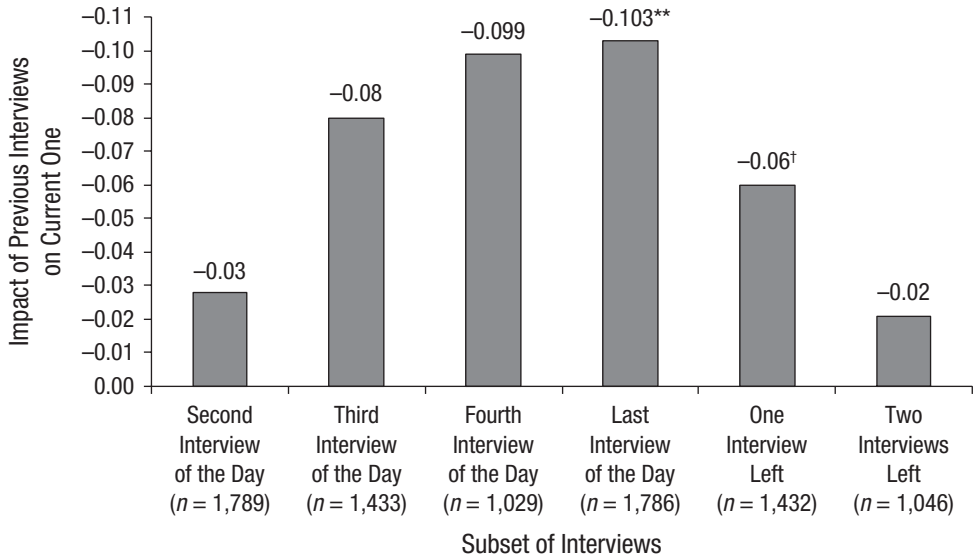

Fig. 2. Impact of the average score given to previous interviewees on the score given to the current applicant as a function of the serial position of the current interview. The graph presents point estimates from Model 4 (see Table 1). Sample sizes varied as a function of the number of interviews that fit the criterion and that had nonmissing information on all variables. Significance of the point estimates is indicated ($^{\dagger}p < .10$; $^{**}p < .01$).

account." The third possibility is that interviewers themselves do not engage in narrow bracketing but believe that people evaluating their performance do, and thus avoid unrepresentative subsets in an attempt to please their audience (as suggested by work on accountability; for a review, see Lerner & Tetlock, 1999).

These mechanisms are not mutually exclusive, and they may coexist. Future research could examine their relative importance in narrow bracketing and, perhaps more important, establish additional consequences of such psychological processes in everyday judgments.

Declaration of Conflicting Interests

The authors declared that they had no conflicts of interest with respect to their authorship or the publication of this article.

Supplemental Material

Additional supporting information may be found at http://pss.sagepub.com/content/by/supplemental-data

Notes

1. Applications for the 2004–2005 academic year were missing from our records.
2. A referee noted that when samples are small, regressions in which a predictor is the average of multiple lags of the dependent variable can lead to spurious negative correlations. To assess if this was a problem in our data, we created 100 mock data sets randomly resorting the interviews by each interviewer across days and then estimated the regressions reported in Table 1 (recomputing mock daily averages). The point estimate of interest was on average −0.002, about 1/50th the size of the estimate in the real data set (i.e., −0.088). Thus, this concern regarding spurious correlations is not a problem in our data set.
3. We also estimated ordered probit and logit regressions. For Model 4, the key point estimates were −0.155, $p = .006$, and −0.250, $p = .014$, respectively. Note that we restricted the sample for all models reported to the third interview onward, believing that only after a few interviews would narrow bracketing have an impact. When we included all interviews, point estimates for Models 1 through 3 were −0.089, −0.079, and −0.076, ps $< .01$, and the point estimate for Model 4 was −0.057, $p < .05$.
4. For ease of exposition, we truncated the x-axis at averages of 2 (lower end) and 4 (upper end) because very few average previous scores (< 5%) were below or above those values. The Supplemental Material includes a table with all values.
5. Point estimates for the subscores were as follows—communication skills: $b = 0.286$; drive: $b = 0.262$; ability to work in teams: $b = 0.168$; accomplishment: $b = 0.247$; and interest in the school: $b = 0.120$. The standard deviations for these subscores were 0.79, 0.71, 0.72, 0.71, and 0.68, respectively. The impact of 1 standard deviation of the communication subscore on overall score, then, was 0.226 (0.286 × 0.79). Dividing 0.075, the impact of the average previous score, by this value yielded 0.33, which means that an increase of 1 standard deviation in the average previous score had an impact equivalent to an increase of 0.33 standard deviation in the communication score. Other calculations were analogous.

References

Brenner, L., Griffin, D., & Koehler, D. J. (2005). Modeling patterns of probability calibration with random support theory: Diagnosing case-based judgment. *Organizational Behavior and Human Decision Processes*, *97*, 64–81.

Gilovich, T., Vallone, R., & Tversky, A. (1985). The hot hand in basketball: On the misperception of random sequences. *Cognitive Psychology*, *17*, 295–314.

Griffin, D., & Tversky, A. (1992). The weighing of evidence and the determinants of confidence. *Cognitive Psychology*, *24*, 411–435.

Higgins, T. E., Rholes, W. S., & Jones, C. R. (1977). Category accessibility and impression formation. *Journal of Experimental Social Psychology*, *13*, 141–154.

Larrick, R. P., & Wu, G. (2007). Claiming a large slice of a small pie: Asymmetric disconfirmation in negotiation. *Journal of Personality and Social Psychology*, *93*, 212–233. doi:10.1037/0022-3514.93.2.212

Lerner, J., & Tetlock, P. (1999). Accounting for the effects of accountability. *Psychological Bulletin*, *125*, 255–275.

Massey, C., & Wu, G. (2005). Detecting regime shifts: The causes of under- and overreaction. *Management Science*, *51*, 932–947.

Read, D., Loewenstein, G. F., & Rabin, M. (1999). Choice bracketing. *Journal of Risk and Uncertainty*, *19*, 171–197.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*, 1359–1366.

Srull, T. K., & Wyer, R. S. (1979). Role of category accessibility in the interpretation of information about persons: Some determinants and implications. *Journal of Personality and Social Psychology*, *37*, 1660–1672. doi:10.1037//0022-3514.37.10.1660

Thaler, R. H. (1985). Mental accounting and consumer choice. *Marketing Science*, *4*, 199–214.

Thaler, R. H. (1999). Mental accounting matters. *Journal of Behavioral Decision Making*, *12*, 183–206.

Thompson, E. P., Roman, R. J., Moskowitz, G. B., Chaiken, S., & Bargh, J. A. (1994). Accuracy motivation attenuates covert priming: The systematic reprocessing of social information. *Journal of Personality and Social Psychology*, *66*, 474–489.

Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin*, *76*, 105–110.

Wheeler, S. C., & Petty, R. E. (2001). The effects of stereotype activation on behavior: A review of possible mechanisms. *Psychological Bulletin*, *127*, 797–826.

Winter, L., & Uleman, J. S. (1984). When are social judgments made? Evidence for the spontaneousness of trait inferences. *Journal of Personality and Social Psychology*, *47*, 237–252.