Judgment and Decision Making, Vol. 8, No. 5, September 2013, pp. 512–520

Belief in the unstructured interview: The persistence of an illusion

Jason Dana* Robyn Dawes† Nathanial Peterson†

Abstract

Unstructured interviews are a ubiquitous tool for making screening decisions despite a vast literature suggesting that they have little validity. We sought to establish reasons why people might persist in the illusion that unstructured interviews are valid and what features about them actually lead to poor predictive accuracy. In three studies, we investigated the propensity for "sensemaking" - the ability for interviewers to make sense of virtually anything the interviewee says—and "dilution"—the tendency for available but non-diagnostic information to weaken the predictive value of quality information. In Study 1, participants predicted two fellow students' semester GPAs from valid background information like prior GPA and, for one of them, an unstructured interview. In one condition, the interview was essentially nonsense in that the interviewee was actually answering questions using a *random* response system. Consistent with sensemaking, participants formed interview impressions just as confidently after getting random responses as they did after real responses. Consistent with dilution, interviews actually led participants to make worse predictions. Study 2 showed that watching a random interview, rather than personally conducting it, did little to mitigate sensemaking. Study 3 showed that participants believe unstructured interviews will help accuracy, so much so that they would rather have random interviews than no interview. People form confident impressions even interviews are defined to be invalid, like our random interview, and these impressions can interfere with the use of valid information. Our simple recommendation for those making screening decisions is not to use them.

Keywords: unstructured interview, random interview, clinical judment, actuarial judgment.

1 Introduction

In 1979, an act of legislature suddenly forced the University of Texas Medical School at Houston to admit 50 more applicants late in the admissions season. The additional applicants were initially rejected for admission, based largely on impressions from *unstructured interviews* in which each interviewer could ask different questions of different applicants in whatever way he or she saw fit. Apparently, the expense of having faculty interview every applicant was wasted: at the conclusion of medical training and one postgraduate year, there were no meaningful differences between the initially rejected and initially accepted groups of students in terms of attrition, academic performance, clinical performance, or honors earned (Devaul et al., 1987). Several large-scale field studies have provided similar examples of the embarrassingly poor validity of unstructured interviews for screening decisions (e.g., Bloom & Brundage, 1947; Milstein, Wilkinson, Burrow, & Kessen, 1981; Carroll, Wiener, Coates, Galegher, & Alibrio, 1982). More systematic reviews in the area of employment decisions likewise show that unstructured interviews are poor predictors of job performance, with structured interviews faring somewhat better (Wiesner & Cronshaw, 1988; Wright, Lichtenfels, & Pursell, 1989; Huffcutt & Arthur, 1994; McDaniel, Whetzel, Schmidt, & Maurer, 1994).

Despite the evidence, unstructured interviews remain a ubiquitous and even predominant tool for many screening decisions. Studies of human resource executives suggest that they believe more in the validity of unstructured interviews than other screening methods, even when they are aware that the evidence suggests that structured assessment is superior (Highhouse, 2008). Academics, though not professional interviewers, may decide to accept graduate students or hire faculty based on an informal 20 minute chat, countermanding substantial aggregated and/or statisticized data comparing the candidate to others (test scores and GPAs in the case of students, C.V.'s in the case of faculty). Recently, Wake Forest University stopped requiring standardized tests for undergraduate admissions, moving to a system in which every applicant is eligible for an unstructured interview that figures into the admissions decision in a "holistic", non-numeric manner (Highhouse & Kostek, 2013).

Since Meehl's (1954) seminal book on the clinical-statistical controversy for making predictions, a large literature studying human vs. statistical prediction has shown that the statistical method is nearly always equal

Dawes was thankful for support from NSF Grant SES-0136259. His unfortunate death in 2010 prevented him from seeing the present version of this article.

*Yale University, 135 Prospect St., New Haven CT 06511, Email: jason.dana@yale.edu.

†Carnegie Mellon University.

or superior to clinical judgment (see Grove, Zald, Lebow, Snitz, & Nelson, 2000). Further, many common objections to the interpretation of this evidence have been thoroughly discussed and refuted (see, e.g., Dawes, Faust, & Meehl, 1989; Grove & Meehl, 1996). We do not seek to rehash these debates or provide merely another example of clinical judges failing to outperform a statistical rule. Rather, we seek to establish some reasons why people might persist in the illusion that unstructured interviews are valid and why they can harm predictions. Specifically, we explore how the inevitably noisy signals in an interview dilute the decision maker's potential use of valid information and how interviewers can form falsely coherent impressions from virtually anything the interviewee says or does.

Extending the literature in these ways is important for at least two reasons. First, it is reasonable to think that, while unstructured interviews are not particularly predictive, they will not *hurt* accuracy. At least, we are aware of no prior evidence that unstructured interviews decrease accuracy, e.g., by way of studying the same decision makers with and without access to interviews. This point becomes important because the issue is often raised that the interview conveys benefits beyond its predictive validity. For example, candidates who go through an interview process may have an increased sense of commitment and more likelihood to accept an offer. Thus, if one can get a benefit from conducting unstructured interviews and the interviews do not make one's judgment *worse*, it would seem riskless to use them for certain purposes. Indeed, this point was made without rebuttal in a discussion of the Wake Forest decision on the Society for Judgment and Decision Making's mailing list (2008–9, see the "search" at http://www.sjdm.org). We will provide evidence that exposure to unstructured interviews can indeed harm judgment.

Second, many have the feeling that the unstructured interview is the best way to uncover important information that is special to a candidate. Particularly, candidates may possess some personality traits at the extremes of the distribution that might make them ideal or unsuitable. These unusual cues, akin to what Meehl (1954) called "broken legs", could immediately remove a candidate from consideration or catapult a candidate ahead of others. Indeed, even a structured interview might not help for this purpose if the interviewer does not have in mind what this broken leg factor would be ahead of time. The logic of broken leg cues was addressed by Meehl, who pointed out that, if people were actually good at spotting broken legs that statistical rules miss, then they would be more accurate than statistical rules. This argument, however, has an important flaw in that it assumes that all errors and successes are equally important. If an interviewer is especially concerned with making some kinds of errors, such as missed broken legs, unstructured interviews could, in theory, be highly valuable in avoiding the most important errors, which may outweigh making a few more errors on the more mundane cases.

Basic psychological research, however, gives us reason to believe that unstructured interviews can harm judgment and reason to doubt that interviewers will be sufficiently adept at spotting special information, and not false alarms, about a candidate.

1.1 Can interviews hurt?

Access to an interview could hurt predictive accuracy because exposure to non-diagnostic information is known to dilute valuable information. Unstructured interviews expose interviewers to so many casual observations about the interviewee that have little or unknown diagnosticity that interviewers cannot help but get more information than they can use and thus, they must ignore some cues. Research on the "dilution effect" (e.g., Nisbett, Zukier, & Lemley, 1981; Zukier, 1982; Peters & Rothbart, 2000) shows that rather than just being ignored, extraneous information reduces reliance on good information. It is perhaps no coincidence that the stimuli for the earliest dilution effect studies, which included ample material judged non-diagnostic by study participants, came from interview snippets (Nisbett, Zukier, & Lemley, 1981). Because making good social judgments often requires ignoring information and relying on simple rules, cognitive traits that might normally be construed as positive, such as complexity of thought and need for cognition, can actually be detrimental to accuracy (Ruscio, 2000). In the presence of quality cues, most of the interview could serve as a distraction.

1.2 Can interviewers reliably extract special information from unstructured interviews?

Although too much irrelevant information dilutes the prediction process, it can also lead to unwarranted confidence due to sensemaking. People seek to impose order on events, so much so that they often see patterns in random sequences (Gilovich, 1991). As such, even the noisiest interview data are readily translated into a good story (see Dawes, 2001, Chapter 7) about the interviewee. Just as one can, post hoc, fit a "significant" statistical model to pure noise, interviewers have too many degrees of freedom to build a coherent story of interviewees' responses. If the interviewee gives a response that is inconsistent with the interviewer's impression, the interviewer can dynamically reformulate that impression, perhaps asking follow up questions until hearing a set of responses that confirm an impression. Without structure,

interviewers may not ask questions intended to disconfirm these impressions. because people are inclined to seek information that confirms their hypotheses or avoid what might disconfirm them (Devine, Hirt, & Gehrke 1990; Sanbonmatsu, Posavac, Kardes, & Mantel, 1998).

As an example, someone known to one of the authors was given a panel interview for a potential job. Arriving 5 minutes early, she was immediately called into the room, where the interview went quite succesfully and she was offered the job on the spot. During the postmortem discussion, one of the interviewers was impressed by how well she composed herself after showing up 25 minutes late to the interview! Apparently, she had been misinformed that the time of the interview was 30 minutes after the hour, rather than on the hour as the panel expected, and she remained composed because she did not know she was late. Interestingly, nothing that the panel asked effectively tested this impression that the candidate was unusually composed under (what they believed were) the circumstances—they never learned that she thought she was early. Further, there are many other less flattering impressions than "composed" that could also have explained a lack of concern over being 25 minutes late, including flippant or arrogant.

The ability to sensemake combined with the tendency for biased testing allows unstructured interviewers to feel they understand an interviewee almost regardless of the information they receive. Unfortunately, a feeling of understanding, while reassuring and confidence-inspiring, is neither sufficient nor necessary for making accurate assessments (Trout, 2002). Further, there is empirical evidence that confidence and accuracy are often poorly related in interpersonal prediction contexts (Dunning, Griffin, Milojkovic, & Ross, 1990; Swann & Gill, 1997) and confidence has been shown to increase with information even in situations where accuracy does not (e.g., Andersson, Edman, & Ekman, 2005; Hall, Ariss, & Todorov, 2007). We suggest that people can feel confident in the validity of unstructured interview impressions even if they are worthless.

We experimentally tested the roles of dilution and sensemaking in the context of using unstructured interviews to predict social outcomes. Study participants predicted the semester GPAs of other students based on biographical information including GPA prior to the semester in question and in some cases, an unstructured interview. In some conditions, the interviews were nonsense for the task at hand because the interviewee secretly used a random responding system to answer questions, literally providing random answers to questions that were independent of the interviewee's natural response. Consistent with dilution, participants' GPA predictions were more accurate without the unstructured interview and less accurate than had they simply predicted that semester GPA would be equal to prior GPA, a strong cue that they were given before making predictions. Consistent with sensemaking, participants who unknowingly conducted random interviews were just as likely to indicate in post-interview surveys that they got good information as those who conducted accruate interviews. This is the first evidence we know of that unstructured interviews can be worse than invalid; they can actually decrease accuracy. Yet, while interviews were harmful in this context, even our nonsense interviews promoted a feeling of confidence in the interview impression.

Table 1: Interviewees' prior and obtained GPAs.

	Interviewees							
Study1	1	2	3	4	5	6	7	
Prior GPA	3.32	3.28	3.24	3.23	2.95	2.84	2.81	
Obtained GPA	3.80	3.08	3.71	3.34	2.68	2.69	3.35	
Study 2	1	2	3	4	5	6	7	8
Prior GPA	3.69	3.38	3.29	3.29	3.23	3.05	2.83	2.65
Obtained GPA	3.83	3.80	4.00	2.83	2.65	3.59	3.00	3.31

2 Study 1

To explore whether interviews could dilute judgments and make them worse, we had student participants predict the semester grade point average (GPA) of two other students, one prediction with biographical information (described below) and an interview, the other with just biographical information. To explore whether interviewers sensemake, we developed a random responding system that the interviewees could use during the interview to see whether it would perturb predictive accuracy or subjective confidence in interview impressions.

2.1 Method

2.1.1 Interviewers and interviewees

Interviewers were 76 undergraduate students at Carnegie Mellon University who were recruited through campus advertising and paid for their participation. We employed five Carnegie Mellon undergraduates (two female) as permanent interviewees. The interviewees ranged in age from 18 to 22, and represented multiple races, majors, and class standings. Two of the interviewees worked for two semesters, creating a total of 7 different semester GPAs to be predicted. Their prior cumulative GPAs and GPAs for the semester to be predicted are listed in Table 1.

2.1.2 Procedures

Participants were introduced to a randomly assigned interviewee and asked to conduct a 20 minute interview with the goal of predicting the interviewee's GPA for a given semester. An experimenter remained in the room during the interview to track time and answer any questions about the task. Prior to interviewing, participants were told the interviewee's age, major, class standing, and course schedule for the semester to be predicted. Participants were offered a break 10 minutes into the interview, during which they could formulate more questions to ask.

After the interview, the interviewee was excused and participants made their GPA predictions, which were to be kept confidential from the interviewee. Before making their predictions, participants were given the interviewee's cumulative GPA prior to the target semester and informed that prior cumulative GPA by itself was the best statistical model for predicting GPAs at this institution (Lewis-Rice, 1989). After the GPA prediction, participants answered a brief questionnaire (Table 3) probing whether they got to know the interviewee and whether the interview provided useful information. Finally, 68 participants predicted the semester GPA for another target whom they did not interview using only the target's background information and cumulative GPA prior to the semester in question.

2.1.3 Interview conditions

The structure of the interview varied according to the participant's random assignment to one of three conditions. In the *accurate* condition ($n = 25$), participants could ask only closed-ended questions, i.e., "yes or no" or "this or that" questions. Interviewees answered these questions accurately. The *random* condition ($n = 26$) was similar except that after the midway break, the interviewee secretly responded on a pseudo-random basis. Interviewees noted the first letter in the last two words of each question and classified them as category 1 (letters A through M) or category 2 (N through Z). If both letters belonged to the same category, the interviewee answered yes (or took the first option of a "this" or "that" question) and otherwise answered no. This system tends to equalize the frequency of yes and no answers as follows: Call the proportion of words that the interviewer samples from category 1. A yes answer occurs if both of the last 2 words are category 1, which occurs with probability p^2, or both category 2, which occurs with probability $(1-p)^2$. The total probability of a yes response, $p^2 + (1-p)^2$, is always closer to .5 than p itself. By employing random response, whether an interviewee's response did or did not match the interviewer's expectations or confirm the interviewer's impression was simply a matter of chance.

A lack of a significant difference in accuracy or survey answers between the accurate and random conditions might not reflect sensemaking on the part of the interviewer, but rather the deficient quality of all closed-ended interviews. That is, if closed-ended interviews are too low in quality for this task, any differences between random and accurate interviews might be muted. In that case, we would expect predictions to be better and ratings to be higher if participants could ask questions and demand answers in any way they wanted. To rule out this explanation, we also conducted a *natural* condition ($n = 25$) in which no closed-ended constraint was placed on the interviewer's questions.

2.2 Results

The validity (correlation with actual outcomes) of GPA predictions following interviews ($r = .31$) was indeed significantly lower than the validity of using prior cumulative GPA alone ($r = .65, t_{(73)} = 3.77, p < .05$, **d** $= .43$; Hotelling's method for dependent r with Williams correction), information participants had when making their predictions. As dilution predicted, our unstructured interview did not prove helpful in light of an already strong cue of prior GPA. While worse than using prior cumulative GPA, a validity of .31 compares favorably to that of unstructured employment interviews for predicting job performance (Campion, Palmer, & Campion, 1997). Comparing the success of our interviewers with employment screeners is not totally appropriate—GPAs could be easier to predict becuase GPA is more reliable than measures of job performance or because job screeners do not have information as valid as prior GPA when making decisions. Still, the validities our interviewers were able to obtain provide at least some evidence that they were not completely deficient at the task.

Some of our interviewees were initially concerned that the random interview would break down and be revealed to be nonsense. No such problems occurred and random interviews proceeded much as interviews in our other conditions. Further, random responding did not make interviewers less accurate: Only the validity in the random condition ($r = .42$) was significantly different from zero, while validities in the accurate ($r = .20$) and natural ($r = .29$) conditions were not, though validities from these 3 conditions did not differ significantly from each other. A plausible concern is that random condition participants might have relied more on prior GPA because the interview was bad, thus inflating accuracy in the random condition because prior GPA was a strong predictor. This was not the case; GPA predictions were no more correlated with prior GPAs in the random condition ($r = .54$) than in the accurate ($r = .53$) or natural condition ($r = .67$).

Table 2: Regression analyses of the accuracy of GPA predictions. (Dependent Variable: Predicted GPA.)

	Study 1	Study 2	All	Study 1	Study 2	All
Predictors:	(1)	(2)	(3)	(4)	(5)	(6)
Actual GPA	0.51^{**}	0.19^{*}	0.09	-0.21	22.67	21.07
	(0.08)	(0.09)	(0.016)	(1.72)	(17.42)	(19.63)
Access to interview	0.99^{**}	0.08	0.54^{*}			
	(0.35)	(0.42)	(0.22)			
Actual GPA×interview	-0.28^{*}	-0.02	-0.15^{*}			
	(0.11)	(0.12)	(0.07)			
Q1				-0.37	0.73^{**}	0.42
				(0.56)	(0.27)	(0.25)
Q2				-0.11	-0.91^{**}	-0.64^{*}
				(0.48)	(0.28)	(0.25)
Actual GPA×Q1				0.13	-0.19^{*}	-0.11
				(0.17)	(0.08)	(0.07)
Actual GPA×Q2				0.03	0.26^{**}	0.18^{*}
				(0.15)	(0.08)	(0.07)
Target dummies	No	No	Yes[1]	No	No	Yes[1]
Clustering at subject level	Yes	Yes	Yes	No	No	No

* $p < 0.05$, ** $p < 0.01$.

1. Two interviewees obtained the same GPA when samples are combined, thus dummies representing each interviewee are included.

2. Eight participants who did not make a prediction without an interview are excluded.

Although participants judged the interview to be somewhat informative, GPA predictions were actually less accurate with interviews ($r = .31$) than without them ($r = .61$). Because these correlations involved different judgments by the same participant, we tested the difference using regression with participant random effects, regressing GPA predictions on actual GPA, a dummy $= 1$ if an interview was conducted, and the interview×GPA interaction. The results in column 1 of Table 2 indicate that the interaction term was negative and significant, meaning that predictions were indeed significantly less correlated with outcomes when an interview was performed.

Table 3 shows that the mean agreement with the statements "I am able to infer a lot about this person given the amount of time we spent together" (accurate $= 2.72$, natural $= 2.80$, random $= 2.85$) and "From the interview, I got information that was valuable in making a GPA prediction" (accurate $= 3.00$, natural $= 3.12$, random $= 3.31$) was similar across all conditions, with no significant differences emerging ($F_{(2,73)} = .233$ and 1.714, respectively). While comparisons of accuracy and subjective impressions yielded null results between random and truthful interviews, in both cases the direction was "wrong"—prediction accuracy and impressions of usefulness trended higher for random interviews. Agreement with these questions did not significantly modulate accuracy. Column 4 of Table 2 reports the results of regressing GPA predictions on obtained GPA, questions 1 and 2, and their interactions. Neither question interacted significantly with GPA.

Table 3: Study 1 post prediction questionnaire. Mean agreement with statements on a 4-point Likert scale (1 = disagree, 4 = agree) with standard errors in parentheses.

	Accurate	Random	Natural
I am able to infer a lot about this person given the amount of time we spent together.	2.72 (.68)	2.85 (.47)	2.80 (.58)
From the interview, I got information that was valuable in making a GPA prediction.	3.00 (.65)	3.31 (.55)	3.12 (.60)

2.3 Discussion

Consistent with sensemaking, a random interview did not perturb either GPA predictions or subjective impressions about the quality of the interview or the extent to which they got to know the interviewee. Consistent with dilution, a single, strong cue—past GPA—predicted better than participants themselves, even though they had this information. Further supporting dilution, participants made better predictions without an interview than with one. While participants generally agreed that they got useful information from interviews, interviews significantly impaired accuracy in this environment.

Perhaps one reason that participants felt interviews were useful and made sense of them even when they were random is that they conducted them. The person conducting the interview controls the questions, which could be important to at least the sensemaking part of our results. If participants merely watched the interviews, rather than conducting them, would they be less prone to either or both effects? By having participants watch pre-recorded interviews, we could also directly assess whether they can tell random from accurate by informing them of the possibility that the interview they watched was random and asking them to guess which type they saw.

3 Study 2

Rather than conducting the interview themselves, participants in Study 2 watched a pre-recorded interview that another student had conducted. Because this procedure did not allow participants to ask their own questions, they could be less prone to confirming their own theories of the interviewee and thus less prone to sensemaking. If so, we might expect participants to be able to discern random from accurate interviews.

3.1 Method

3.1.1 Participants and interviewees

Participants were 64 undergraduate students at Carnegie Mellon University who were recruited through campus advertising and paid for their participation. Eight Carnegie Mellon undergraduates (5 female) participated as interviewees and consented to having two interview sessions recorded (one random, one accurate) as stimuli for the study. Interviewees ranged in age from 19 to 21, and again represented multiple races, majors, and class standings. Table 1 lists their prior and obtained GPAs.

3.1.2 Procedures

Procedures were the same as in Study 1, with the following exceptions. Prior to conducting the experimental sessions, we video-recorded 16 interviews (one accurate and one random for each interviewee, natural interviews were not used) conducted similarly to Study 1, except that the random interview was now random responding throughout instead of after the break. Participants were randomly assigned to watch one of the 16 interviews via computer interface and predict the interviewee's GPA for a given semester. Each interview was randomly assigned to four different participants. The post interview question wording was amended slightly (Table 4) to reference the interview that was watched and the Likert-type scale now ranged from 1 to 5 and included a "neither agree nor disagree" point. After the post-interview questionnaire, participants were informed that their interview was randomly drawn from a pool containing half random interviews and asked to guess whether it was random or accurate.

Table 4: Study 2 mean Likert responses (5 = strongly agree) to post experimental questions by condition (standard errors in parentheses).

	Accurate	Random
I am able to infer a lot about this person given the interview I just watched.	3.47 (0.92)	3.47 (1.08)
From the interview I just watched, I got information that was valuable in making a GPA prediction.	3.66 (0.94)	3.75 (0.98)

3.2 Results

GPA predictions were about equally correlated with actual GPAs as they were in Study 1 ($r = .28$). For this sample of interviewees, however, prior GPA was not as predictive of semester GPAs as it was in Study 1 ($r = .37$) and was not significantly more accurate than participant predictions. Thus, this form of dilution was not present in Study 2. Though the procedure in Study 2 is somewhat different, it is informative to combine results with Study 1 to see if, overall, dilution is present, especially considering that the sample of interviewees on which this result somewhat depends is small. Combining both studies, prior GPA alone predicts significantly better than our participants do with interviews ($t_{(137)} = 2.59, p < .05$, $\mathbf{d} = .44$).

Even though participants did not control the course of the interview in Study 2, subjective impressions were again unperturbed by random responding. Table 4 shows that mean agreement with the statements "I am able to infer a lot about this person given the interview I just

Table 5: Frequency of accurate/random guesses by interview type.

	Random	Accurate	Total
Guess Random	13	3	16
Guess Accurate	19	29	48
Total	32	32	64

watched" (accurate $= 3.47$, random $= 3.47$) and "From watching the interview, I got information that was valuable in making a GPA prediction" (accurate $= 3.66$, random $= 3.75$) was again similar across conditions, with agreement in the random condition again being equal or higher. As in Study 1, GPA predictions relied on prior GPA about the same for random ($r = .58$) and accurate ($r = .55$) interviews.

We again tested for interactions between answers to the post-experimental questionnaire and predictive accuracy. As can be seen in column 5 of Table 2, both interactions were significant in Study 2. Interestingly, the coefficients on each question and on each interaction had opposite signs, such that feeling one is able to infer a lot about the interviewee negatively impacted accuracy, while feeling one had gotten good information from the interview positively impacted accuracy. When studies 1 and 2 are combined, shown in column 6 of Table 2, only the interaction between the valuable information question and accuracy remained significant. This result seems somewhat paradoxical: Access to interviews overall decreased accuracy, but given that a participant had access to an interview, greater agreement that one had gotten valuable information from the interview increased accuracy. This result raises the question of whether our finding of poorer accuracy following interviews is driven by a subset of participants who did not feel they got useful information from the interview, but used it anyway. The effects are not so simple, however. For example, looking at only those participants who agreed with the valuable information question (answers of 4 or 5), the validity of predictions was only .29. Thus, there is no simple main effect such that those who felt they got valuable information from the interview were more accurate.

Table 5 tabulates participants' judgments of whether they saw an accurate or random interview across interview type. Participants correctly classified 66% of the interviews, significantly better than chance ($\chi^2_{(1)} = 8.33, p < .01$). This result, however, was largely driven by the participants judging all interviews to be accurate: accurate interviews were nearly always judged to be accurate ($29/32$), and more than half of random interviews were judged accurate ($19/32$). Indeed, the tendency to judge all interviews accurate was significantly stronger than the tendency to be correct (McNemar's test, $\chi^2_{(1)} = 11.63, p < .001$). Thus, while participants have some skill in identifying accurate from random interviews, they also see most interviews as probably being accurate, indicating some degree of sensemaking. Whether participants were accurate in this judgment, whether participants judged their interview to be accurate, and whether participants correctly judged their interviews to be accurate all did not interact with accuracy of GPA predictions (all $p > .30$).

Although participants who merely watched interviews were still prone to sensemaking, their predictions were not more accurate without an interview, inconsistent with dilution (see column 2 of Table 2). Dilution did not hold for these participants, however, largely because no-interview predictions, which were not handled differently in this study, were much less accurate ($r = .26$) than in Study 1, while predictions following all interviews were about as accurate as in Study 1 ($r = .28$). Of course, while interviews did not make predictions worse, they also did not make them significantly better. Our two studies thus fail to indicate any incremental validity from interviews, and Study 1 suggests a decrement in validity. At best, one can say that watching an interview did not hurt but conducting one did. While the procedures are different across the studies, it is again informative to combine the data and repeat our test of predictive validity with and without an interview. Column 3 of Table 2 shows that the interview$\times$GPA interaction is negative and significant; thus, interviews are overall negatively associated with accuracy.

3.3 Discussion

Watching interviews did little to mitigate sensemaking; participants' predictive accuracy and subjective impressions were similar after watching random and accurate interviews, and they were more likely to see interviews as accurate whether they were or not. One objection to our interpretation of Studies 1 and 2 is the presence of experimental demand to use interviews. Because we took the trouble of having participants conduct or watch interviews for the majority of the study's duration, it is not unreasonable to assume that participants felt they should use the interview, regardless of their feelings about its validity. Of course, such implicit demands are also present in real-world settings in which one is forced to conduct an interview for screening purposes. Still, one may wonder whether participants believed that interviews aided accuracy, a question we explore in Study 3.

Table 6: Dominance matrix in which cell frequencies are the number of participants who ranked the column method better than the row method.

	Natural	Accurate	Random	No interview	Total
Natural	–	36	12	13	61
Accurate	128	–	28	22	178
Random	153	136	–	68	357
No interview	152	142	96	–	390

4 Study 3

4.1 Method

One hundred sixty nine Carnegie Mellon University students completed this task as part of a longer session. Participants were given descriptions of the methods and conditions used in Study 1 (except that the random condition was full random as in Study 2), including the information that participants were given prior GPA and then asked to predict a student's GPA from a given semester. Participants in Study 3 were then asked to rank the interview types (including no interview) in terms which they would like to have to make their predictions as accurately as possible. That is, they were essentially asked about the incremental validity of each type of interview.

4.2 Results

The modal accuracy rankings for first through last place were natural interview first, followed by accurate, random, and no interview, respectively, making the prediction type for which participants were the most accuate in Study 1 the least favored. This ranking was also the single most common, chosen by 57 (33%) of our participants. No participant ranked the natural condition last, while 56% of participants ranked no interview last. The dominance matrix in Table 6 depicts all aggregate pairwise preferences by reporting how many participants ranked the interview type in the column over the type in the row. Even random interviews, which by definition contain misleading information, were preferred to no interview by 96 participants (57%). By ranking the random interview ahead of no interview, a simple majority of our participants showed that they did not anticipate a dilution effect: Apparently, they believed that random interviews contained *some* useful information that all of the useless information would not drown out. Thus, while interviews do not help predict one's GPA, and may be harmful, our participants believe that any interview is better than no interview, even in the presence of excellent biographical information like prior GPA.

5 Discussion

We set out to examine whether unstructured interviews could harm predictive accuracy and whether interviewers would believe they garnered useful information from the interview regardless of its quality. Consistent with dilution, Study 1 showed that participants were better at predicting other students' GPAs when they were not given access to an unstructured interview in addition to background information. Further, participants predicted worse than if they had used prior GPA alone, information they were given before making their predictions. Consistent with sensemaking, participants were just as able to make coherent impressions when the interviewee responded randomly, both in terms of the accuracy of their predictions and their confidence in their subjective impressions. Study 2 showed that even when watching rather than conducting an interview, participants were still somewhat prone to sensemaking. Finally, Study 3 showed that participants believe that interviews will help in this context, so much so that they rate random interviews as being more helpful than no interview, which was, in fact, the best way to make predictions in Study 1 and as good as other methods in Study 2.

Our findings suggest a rethinking of the meaning of interview validity. The validity of predictions made by interviewers or by numerically incorporating interviews into a model is uninformative unless it can be directly compared to predictions made by the same methods without an interview. On its face, the validity of our participants' predictions following unstructured interviews looks respectable ($r = .31$ in Study 1 and $r = .28$ in Study 2), yet these same participants were able to predict better when they did not have access to an interview, and could have predicted better still if they just used prior GPA. It may be the case that for many screening decisions, there are one or two cues that are very important and could be garnered from nearly any interview (or without one), and that these cues predict better by themselves than the clinical judges who have access to them. The substantial literature on interviews for employment screening, which already indicates that unstructured interviews are not particularly good, may thus even be overstating the validity of unstructured interviews. Our evidence is experimental and compares the same judges with and without access to an interview. To our knowledge, there is little prior evidence of this kind.

In addition to the vast evidence suggesting that unstructured interviews do not provide incremental validity, we provide direct evidence that they can harm accuracy. Because of dilution, this finding should be especially applicable when interviewers already have valid biographical information at their disposal and try to use the unstructured interview to augment it. Because of sensemak-

ing, interviewers are likely to feel they are getting useful information from unstructured interviews, even when they are useless. Because of both of these powerful cognitive biases, interviewers probably over-value unstructured interviews. Our simple recommendation for those who make screening decisions is not to use them.

References

Andersson, P., Edman, J., & Ekman, M. (2005). Predicting the world cup 2002 in soccer: Performance and confidence of experts and non-experts. *International Journal of Forecasting, 21*, 565–576.

Bloom, R. F., & Brundage, E. G. (1947). Predictions of success in elementary school for enlisted personnel. In D. B. Stuit (Ed.), *Personnel research and test development in the Naval Bureau of Personnel*, pp. 233–261.. Princeton: Princeton University Press.

Campion, M. A., Palmer, D. K., & Campion, J. E. (1997). A review of structure in the selection interview. *Personnel Psychology, 50*, 655–702.

Carroll, J. S., Wiener, R. L., Coates, D., Galegher, J., & Alibrio, J. J. (1982). Evaluation, diagnosis, and prediction in parole decision making. *Law & Society Review, 17*, 199–228.

Dawes, R. M. (2001). *Everyday irrationality: How pseudoscientists, lunatics, and the rest of us systematically fail to think rationally*. Boulder, Colorado: Westview Press.

DeVaul, R., Jervey, F., Chappell, J., Caver, P., Short, B., & O'Keefe, S. (1987). Medical school performance of initially rejected students. *Journal of the American Medical Association, 257*, 47–51.

Devine, P., Hirt, E., & Gehrke, E. (1990). Diagnostic and confirmation strategies in trait hypothesis testing. *Journal of Personality and Social Psychology, 58*, 952–963.

Garfinkel, H. (1967). Common sense knowledge of social structures: The documentary method of interpretation in lay and professional fact finding. In H. Garfinkel (Ed.), *Studies in ethnomethodology*, pp. 76–103. Englewood Cliffs, NJ: Prentice-Hall.

Gilovich, T. (1991). *How we know what isn't so: The fallibility of human reason in everyday life.* New York: Free Press.

Hall, C. C., Ariss, L., & Todorov, A. (2007). The illusion of knowledge: When more information reduces accuracy and increases confidence. *Organizational Behavior and Human Decision Processes, 103*, 277–290.

Highhouse, S. (2008). Stubborn reliance on intuition and subjectivity in employee selection. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 1*, 333–342.

Highhouse, S., & Kostek, J. A. (2013). Holistic assessment for selection and placement. In K. F. Geisinger, B. A. Bracken, J. F. Carlson, J-I. C. Hansen, N. R. Kuncel, S. P. Reise, & M. C. Rodriguez (Eds.), *APA Handbook of Testing and Assessment in Psychology. Vol. 1: Test theory and testing and assessment in industrial and organizational psychology*, pp. 565–577. Washington, DC: American Psychological Association.

Lewis-Rice, M. (1989). Marketing post-secondary educational programs with implications of higher education administration. Doctoral dissertation, School of Urban and Public Affairs, Carnegie Mellon University.

Meehl, P. E. (1954). *Clinical versus statistical prediction; a theoretical analysis and review of the evidence*. Minneapolis: University of Minnesota Press.

Milstein, R. M., Wilkinson, L., Burrow, G. N., & Kessen, W. (1981). Admission decisions and performance during medical school. *Journal of Medical Education, 56*, 77–82.

Nisbett, R., Zukier, H., & Lemley, R. (1981). The dilution effect: Nondiagnostic information weakens the implications of diagnostic information. *Cognitive Psychology, 13*, 248–277.

Ruscio, J. (2000). The role of complex thought in clinical prediction: Social accountability and the need for cognition. *Journal of Consulting and Clinical Psychology, 68*, 145–154.

Sanbonmatsu, D., Posavac, S., Kardes, F., & Mantel, S. (1998). Selective Hypothesis Testing, *Psychonomic Bulletin & Review, 5*, 197–220.

Swann, W. B., & Gill, M. J. (1997). Confidence and accuracy in person perception: Do we know what we think we know about our relationship partners? *Journal of Personality and Social Psychology, 73*, 747–757.

Trout, J. D. (2002). Scientific explanation and the sense of understanding. *Philosophy of Science, 69*, 212–233.