Enhancing Consistency of Maximal Responding in Behavior Description Interviews: An Exploration of Priming and Response Length

Allen I. Huffcut[1], Satoris S. Howes[2], Dianne D. Murphy[1], and Sara A. Murphy[3]

1. University of Wisconsin - Green Bay
2. Oregon State University
3. University of Winnipeg, Canada

ABSTRACT

In a behavior description interview (BDI), candidates are asked to describe past experiences that demonstrate skills and abilities important for the position (Janz, 1982). A recent study by Huffcutt et al. (2020) found that only around half of participants (48.1%) describe an experience reflecting maximal performance capability. Random mixing of maximal capability with day-to-day typical performance tendencies is problematic psychometrically because candidates are not all providing comparable information and top candidates could be overlooked. Given notable methodological concerns with Huffcutt et al.'s approach, our first purpose was to provide empirical confirmation that maximal responding in BDIs is, in fact, inconsistent. Our estimate of the proportion of maximal responding was even lower (41.3%), further amplifying concerns when assessment of maximal performance capability is desired (e.g., for many professional positions). The second purpose was to investigate two factors that could increase the consistency of maximal responding: rewording the main BDI question to focus directly on absolute top-end experiences (i.e., priming) and longer response length. Both were found to have significant effects. A number of directions for future research were identified, which, along with these findings, could help researchers move closer to the long-term goal of uniform description of experiences that reflect each candidate's maximal capability (or typical tendencies if so desired).

KEYWORDS

job applicant interviews, behavior description interviews, STAR interviews, typical performance, maximal performance

Behavior description interviews (BDIs), where candidates are asked to relay past experiences that demonstrate skills and abilities important for the position to which they are applying, have emerged as a premier assessment technique.[1] They are predictive of job performance (Culbertson et al, 2017), have minimal impact on protected groups (Huffcutt & Roth, 1998), and are generally well liked by both interviewers and candidates (Bragger et al., 2016). Indeed, it is difficult to find other selection approaches that have such a highly desirable combination of features.

That said, two psychometric concerns have been raised. First, research suggests that half, or fewer, of interviewees recount an actual past experience by themselves (Bangerter et al., 2014; Brosy et al., 2020). Fortunately, careful follow-up questioning (probing) appears to help generate real experiences much more consistently (Brosy et al.). Second, roughly half of candidates describe an experience that reflects their typical (day-to-day) tendencies, whereas the other half describe an experience that portrays their more maximal, top-end capability (Huffcutt et al., 2020). Typical and maximal work behavior reflect very different aspects of job performance and have only a modest intercorrelation. Randomly mixing them increases the possibility of overlooking top candidates and may even contribute to biased ratings.

Despite the potential importance and implications of Huffcutt et al.'s (2020) findings, several major methodological concerns make their results somewhat tentative.

Corresponding author:
Allen I. Huffcut
Author Email: huffcuta@uwgb.edu

1 A common format for responding to BDI questions is often called STAR in business and industry (situation, task, action, result; see Birt, 2022).

First, they relied upon subjective impressions of the participants themselves to determine whether their handling of a difficult customer reflected typical or maximal capability. Second, they formed two very different groups depending on whether participants recalled multiple or only one past experience, which received different follow-up questioning and then combined them to determine the overall proportion of maximal (and typical) responding. Finally, their sample consisted primarily of undergraduate college students working part time in general retail. We expound upon these limitations in more detail later.

The first purpose of the current investigation is to derive a more robust and generalizable estimate of the relative proportion of maximal responding in BDIs, which we then compare to that of Huffcutt et al. (2020). There are situations where assessment of typical performance may be desired, but for many professional positions, the large standard deviation in performance often makes capturing maximal performance advantageous. As such, we focus on maximal capability in this investigation and leave treatment of typical performance for future research. Our approach affords two advantages. First, our sample comprises working adults (in customer / client service positions), who span a surprisingly wide range of positions (e.g., waiter to doctor), ages, ethnicity/nationality, and geography (e.g., from five continents). Second, we employ a more objective and standardized methodology that does not rely on subjective impressions of the participants.

The second purpose is to explore how priming interviewees affects the prevalence of maximal responding. If Huffcutt et al.'s (2020) findings are even ball-park representative, it does not matter if an organization strives to assess typical or maximal capability as BDIs seem to be providing a somewhat random mixture of both. We focus on increasing maximal responding by rewording (i.e., priming) the main BDI question to emphasize identification of an experience that portrays their *absolute best* handling of a difficult customer. Theoretically, priming appears to activate representations in memory consistent with the nature of the priming and has been found to exert significant influence on memory recall and behavior (Aarts et al., 2007).

The third purpose is to explore the association between response length (i.e., number of words used) and the prevalence of maximal responding. Bangerter et al. (2014) found a significant, positive correlation between response length and description of a real past experience (e.g., as opposed to expressing a generality), thereby highlighting its potential for other types of BDI analyses. It appears that working memory is surprisingly limited (four slots on average; Cowan, 2010) and quite susceptible to distractions. Moreover, past experiences appear to be stored as scattered fragments across the outer surface of the brain (Loftus, 1995), which must be retrieved and integrated back into a coherent story (see Huffcutt & Howes, 2023). Extending BDI response lengths could afford candidates more time to successfully navigate this process, thereby allowing a greater number of fragments to be retrieved and integrated.

The fourth and more supplemental purpose is to introduce a companion approach to monitor the effects of priming and increased response length. Although typical versus maximal responding is clearly a within-person phenomenon (i.e., intraindividual performance; Lievens et al., 2018), changes at this level could impact selection outcomes such as utilization of the BDI rating scale. To illustrate, because maximal experiences should get rated more highly than typical ones, more consistent maximal responding is likely to increase the overall BDI mean across all candidates. If the overall mean gets too high, ratings may condense at the top of the scale, making it more difficult to identify the top candidates. Given its centrality to this investigation, we now explore differences between typical and maximal performance in more detail.

Theoretical Perspectives on Typical Versus Maximal Performance

Typical performance reflects the level of effort one puts into their work when they are not being closely monitored, there are no overt instructions to maximize effort, and the time period is long enough for stable patterns to emerge (Sackett et al., 1988). Conversely, maximal performance occurs when workers are aware they are being evaluated, there are explicit instructions to maximize effort, and the time period is short enough to allow sustained effort. In Sackett et al.'s study of grocery store cashiers, for example, the register system automatically recorded actual scan rates over a 4-week period to assess typical effort. Then, without customers present and while being paid overtime, cashiers scanned carts with 25 specific items under direct instructions to work quickly and while being observed to capture their maximal performance level.

What makes the typical versus maximal distinction particularly important is their relatively low correlation. With the grocery cashiers, for instance, these two aspects of performance correlated only .32 among established employees and .14 with new employees. Similarly, Marcus et al. (2007) found that maximal and typical performance correlated around .25 in a sample of middle managers. As Sackett (2007) noted, the question of "Who are our best employees?" can have a very different answer depending upon which of these two aspects is considered (p. 180).

There is a strong argument for assessing maximal capability in more professional positions where the standard deviation of performance is relatively high. Judiesch et al. (1992) summarized across a number of individual studies and found that the average dollar value for performance at the 85th percentile was 72% higher than that at the 50th percentile (essentially one standard deviation difference; see their Table 2). Applying their findings to a hypothetical

sales position, someone at the 85th percentile would be projected to sell $172,000, which is substantially more in comparison to someone at the 50th percentile who sells $100,000 (see also Schmidt & Hunter, 1983).

However, as noted earlier, there are situations where assessment of typical tendencies may be preferred. In some nonprofessional occupations, there may be little or no opportunity to display maximum capabilities and/or where operating at a maximal level makes little difference. For example, a file clerk who puts away paperwork at maximal speed will probably have dead time later. Further, there may be jobs where the standard deviation of performance (per utility analysis) is relatively small, such as janitorial positions, thereby diluting the benefits of higher maximal capability. Regardless of preference, Huffcutt et al.'s (2020) finding of mixed responding and the methodology that generated it warrant careful scrutiny, which we now do.

Overview of Huffcutt et al. (2020)

To our knowledge, this was the first study to investigate the nature of BDI responding directly (i.e., by evaluating the content of the responses). Previous work compared means of BDIs to other structured interview formats[2] (e.g., Morgeson et al., 2005) and/or assessed correlations with differentiated performance ratings (e.g., Klehe & Latham, 2006). Such approaches, although meaningful, represent more indirect evidence.

Huffcutt et al. (2020) analyzed 109 college students from a Midwestern university who were working at least part time in a retail position (e.g., Walmart, GAP, Forever 21). Their primary BDI question was "Tell me about a specific time when you had to deal with a person that was being difficult." Immediately after giving their response, participants were asked follow-up questions such as how many experiences came to mind; whether memory characteristics such as recency, personal impact, frequency, and/or being retail related influenced their recall; and how effectively they think they handled their difficult person.

Subsequently, they formed two separate groups depending on whether participants recalled only one or multiple experiences after hearing the question. In the single recall group, the experiences participants described were deemed to reflect typical performance if they subjectively indicated that their handling of the difficult customer was "in line with what they normally would do" and maximal if it was "better than they usually would" (p. 458). The problem is that "better than usual" could theoretically include handling that was just modestly above average but nowhere near top end, yet these would still have been viewed as maximal. In the multiple recall group, responses were deemed to reflect maximal performance capability if participants felt that the experience they chose to describe was the best among those recalled. The best experience recalled could still reflect somewhat typical performance capability or perhaps modestly above it, but again nowhere near top end.

A total of 51 out of 109 participants (46.8%) reported that only a single experience came to mind. Among this group, 30 indicated that they handled their difficult person in line with what they normally would do, whereas 16 reported handling them better than they normally would. Among the 58 who recalled multiple experiences (53.2%), 34 reported that the experience they chose to describe was the best one among those recalled, whereas 24 reported that it was not the best one.

In a curious methodological approach, these two seemingly disparate groups were combined together for the final analysis, where a roughly even split was found between typical and maximal responding (51.9 vs. 48.1%). The typical group included the 30 participants who recalled a single experience and reported normal handling, and the 24 participants who recalled multiple experiences but did not report their best one (54 total). The maximal group included the 16 participants who recalled a single experience and reported better than normal handling and the 34 participants who recalled multiple experiences and described their best one (50 total). The difference between the two resulting sample proportions (.519 vs .481) was not significantly different. (Note the total sample size was 108 rather than 109; one participant could not be classified.)

Although Huffcutt et al.'s (2020) findings are problematic from a psychometric perspective, it is important to establish that there are meaningful practical effects as well. Doing so is especially important given recent meta-analytic findings that structured interview formats provide a level of criterion-related validity that is on par with the best predictors available (Sackett et al., 2022). We now endeavor to show that there are important practical effects when mixing typical and maximal responding.

Effects of Typical Versus Maximal Responding on Actual Selection

We now develop a thorough and convincing case for why typical versus maximal responding matters at a practical (operational) level in actual selection. Figure 1 shows hypothetical distributions for two associates with mean performance levels that are one standard deviation apart. We display these distributions as being relatively normal, which is intended for the sake of familiarity and convenience rather than to imply normality.

These distributions implicitly assume that there is significant variability in an individual worker's performance over time (i.e., intraindividual), a phenomenon that is reasonable and widely acknowledged in the performance literature (see Barns & Morgeson, 2007; Deadrick & Gardner, 2008). Using handling of difficult customers as a context

2 For instance, the situational interview (Latham et al., 1980).

FIGURE 1.

Maximal Entry Points for Associate 1 (M1) and Associate 2 (M2)

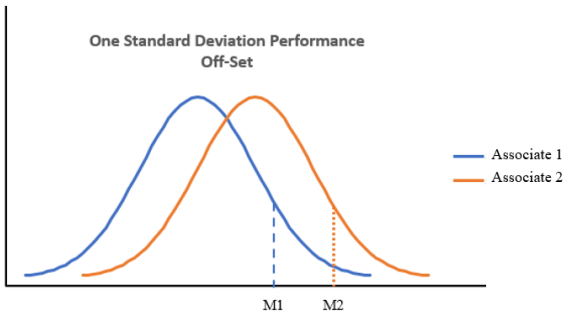
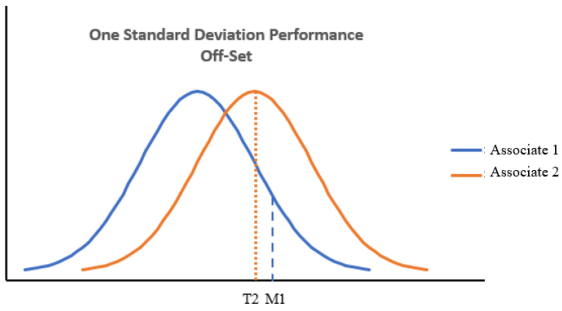


FIGURE 2.

Illustration of Potential Effects Typical Versus Maximal Responding on Selection Outcomes



Note. "T2" denotes modal (typical) performance for Applicant 2, whereas "M1" denotes the beginning of the maximal performance region for Applicant 1.

(from Huffcutt et al., 2020), no associate responds exactly the same in every situation. Contextual influences such as the demographics of the customer (e.g., age, ethnicity), the nature of their complaint, time of day, personal status (e.g., health, stress level), and other factors can all exert influence on how effectively they respond. Over time, a range of effectiveness is likely to form that is unique to each associate, spanning from worst to best handling. Assuming sufficient time in a job, this range should be reasonably stable.

Given the existence of an effectiveness range unique to each person, there should be a region at the top of that range that can reasonably be classified as representing maximal effectiveness for that individual. These regions most likely have an entry point, denoted as M1 and M2 in Figure 1.[3] Note that Associate 2's maximal entry point is noticeably higher than that for Associate 1, and that Associate 1's maximal region barely penetrates Associate 2's region. Because much of Associate 2's maximal region is unique, the implication is that this associate can display a level of performance that Associate 1 is generally unable to reach.

An important and potentially problematic implication thus arises, which is portrayed in Figure 2. Assume that both associates are given a BDI for the same open position. If Associate 1 happens to provide a maximal experience (for them) while Associate 2 provides one that is more reflective of day-to-day responding (i.e., their typical level), Associate 1 would likely be hired even though Associate 2 has higher levels of both typical and maximal performance. Associate 1 may turn out to be a perfectly acceptable employee. However, not only would Associate 2 have had a higher level of performance day-to-day, but as emphasized by Aguinis and O'Boyle (2014), the opportunity to hire someone who has the potential to make an extraordinary (rather than an ordinary) contribution to organizational success is lost (see also O'Boyle & Aguinis, 2012).

Impact of Enhanced Maximal Responding on BDI Scale Utilization

Increasing the consistency of maximal responding should increase subsequent interviewer ratings as a whole, as more maximal experiences generally receive higher ratings. The increase in the overall mean rating across candidates could eventually constrain variability at the top end of the BDI rating scale. Less variability could, in turn, reduce the degree to which truly top candidates stand out, presenting a threat to BDI efficacy as a selection instrument. Modifications to the rating scale may then become necessary, such as increasing the number of scale points or even switching to a different format (e.g., checklist-type scoring).

Unfortunately, neither the intraindividual performance literature nor the BDI literature offered any guidance for identifying the maximal entry point. After careful deliberation, we decided to assess the proportion of candidates who receive a mean rating of 4.0 or higher on our five-point scale as a means to monitor the effects of enhanced maximal responding. In our rating scale, 3 reflected appropriate but not overly extensive actions to resolve the customer difficulty, 4 involved a concerted effort which could have included multiple meaningful actions, and 5 indicated that the associate went way out of their way to satisfy the customer, including making every possible effort. The 4.0 point appears to represent an operational transition between adequate (e.g., "not overly extensive") and much more substantive (e.g., "concerted") performance. We fully acknowledge the subjective nature of choosing this point, and that it could be different depending on the BDI rating scale developed.

3 The degree to which there is an identifiable transition point from typical to maximal performance is interesting scientifically and should be investigated in future research. We assume such a point exists in order to develop the theoretical arguments in this section.

Statement of Study Hypotheses

Huffcutt et al.'s (2020) findings are troubling because they suggest that BDIs do not consistently assess maximal performance capabilities. Unfortunately, there is reason to believe that their estimate of 48% maximal responding may actually be too high. As noted earlier, they formed two disparate groups (single vs. multiple recall of experiences) and then effectively allowed the participants themselves to determine if the experience described was typical or maximal. The wording of their follow-up questioning was such that experiences reflecting modestly better than usual handling of a difficult customer could still be classified as maximal. Using classification methodology that is more standardized and does not rely on subjective impressions of the participants should be more stringent and could suggest an even lower rate of maximal assessment. We hypothesize:

> **Hypothesis 1**: The proportion of maximal responding in our study will be significantly lower than the .481 value found by Huffcutt et al. (2020).

A casual review of the literature suggests that wording BDI questions in a manner that allows for typical responding is common. For example, asking candidates to describe a time when they handled a difficult customer, solved a difficult problem, led a group to a successful outcome, or motivated someone to work harder does not necessarily prime respondents to think of their absolute top-end capability. Consider an IT position where candidates are asked about a time when they solved a difficult computer issue. The problem is that a perfectly ordinary solution (e.g., rebooting) might be sufficient to resolve it. As such, the search for relevant experiences in long-term memory is likely to be fairly broad, and either a typical or maximal experience could result.

Conversely, rewording a BDI question to focus directly on top-end experiences (that maximally portray capabilities) has the potential to narrow the parameters of the long-term memory search process and focus it more strongly on these types of experiences. With a much stronger representation of high-end experiences, the probability of a maximal experience being reported should go up. We hypothesize:

> **Hypothesis 2**: Mean ratings for our primed question (BDI-2) will be significantly higher than mean ratings for the main study question worded in a standard manner (BDI-1).

> **Hypothesis 3**: The proportion of responses reaching a rating of 4.0 or higher will be significantly higher for the primed question (BDI-2) relative to the main study question (BDI-1).

Given the four-slot limitation in working memory and that past experiences are stored as scattered fragments across the outer surface of the brain that must be located and reintegrated, some have argued that the "cognitive load" of BDIs is just too high for a number of candidates (e.g., Brosy et al., 2020; Huffcutt et al., 2020). Extending response lengths could allow candidates more processing time and retrieval of a greater number of fragments. We hypothesize:

> **Hypothesis 4**. Response length (word count) will correlate positively and significantly with mean interviewer ratings for the traditionally worded question (BDI-1), such that longer responses are associated with higher ratings and more maximal responding.

> **Hypothesis 5**. Response length (word count) will correlate positively and significantly with mean interviewer ratings for the primed question (BDI-2), such that longer responses are associated with higher ratings.

METHOD

Participants

Participants were recruited via the online survey posting site Prolific (www.prolific.co). The screening criteria required that participants: (a) be at least 18 years of age, (b) work currently in a direct customer or client contact position where they could take direct action to resolve unhappy individuals, and (c) be in that position for a minimum of 1 year. Participants were eliminated if they failed to complete the survey or completed the study too quickly (see Roulin & Krings, 2016). Further, we included three randomly interspersed attention checks, and participants that failed one or more of them were removed from the study. There was no constraint on how long participants could take to complete the study, and completion time ranged from 3 to 560 minutes with a mean of 37.6 and a median of 31.[4]

A total of 208 individuals completed the study, but three were dropped immediately because they did not answer one or more BDI questions. Among the remaining 205, ages ranged from 18 to 59 years of age (M = 29, Median = 28, SD = 7.7), and 108 reported their gender as male (52.7%), 93 as female (45.4%), and four as nonbinary (2.0%). Participants were primarily White (57.6%), Hispanic-Latino (25.4%), Black (4.4%), and Asian (4.4%). Job positions were surprisingly diverse, thereby enhancing generalizability, and included positions such as server, cashier, customer service rep, bartender, accountant, dentist, doctor,

4 The 560 value is clearly an outlier, as the next highest person took 88 minutes. English was not the native language for both of these participants. At the suggestion of a reviewer, we correlated completion time with mean ratings for BDI-1 and found a very minimal association (r = .07, *ns*).

clinical psychologist, engineer, human resource manager, software engineer, supervisor/manager, teacher, lab technician, and chef.

Procedure

Participants first reviewed and endorsed the informed consent form and then provided demographic and work information. Subsequently, they were presented with the three BDI questions via a video recording by a person portraying a selection professional. Prior to presenting the first question, he introduced the format and explained that once he finished saying each question, participants would type their responses into the corresponding textbox. Although each question was first presented verbally, it was also repeated in written form above the text box provided for response entry.

BDI Questions

The first BDI question, which served as the traditionally worded question (BDI-1), was "Tell me about a specific time you had to deal with a customer who was being difficult." This question was almost identical to the primary question asked in Huffcutt et al. (2020), which was "Think about a time you had to deal with a person that was being difficult." We believe this wording is very consistent with how BDI questions appear to be worded in much of the interview literature.

The remaining two BDI questions were designed to capture the endpoints of each participant's effectiveness (high and low), which allowed us to compute their range. Specifically, the second question (BDI-2), which captured the high endpoint, was "Think for a moment about all the difficult customers you have dealt with during your time in this job. Try to identify one specific instance where you really handled that person in a highly effective manner. In other words, tell me about a particular customer experience where you were at your absolute best." The third question (BDI-3), which captured the low endpoint, was "Now let's go the other way. Think again about all the difficult customers you have had to deal with during your time in this job, but this time I want you to pinpoint one where you really blew it and messed things up."[5]

BDI Response Ratings

Six research assistants served as raters for the BDI responses. Raters were split randomly into two teams of three. Each team rated all participants (with each rater randomly assigned to a separate question, BDI-1, BDI-2, or BDI-3, for each participant), thereby providing two sets of ratings that were averaged to form an overall rating. To minimize rater effects, raters were blind to study hypotheses and to which question they were assessing for any given participant.

The same five-point rating scale was used for all three BDI questions. Behavioral anchors for scale points 3–5 were provided earlier. The first two points were: 1: reacted in kind to customer (e.g., got mad if the customer did so)/ actions were inappropriate/totally misread the situation, and 2: took no real action or took some action but it was the bare minimum/actions were acceptable but superficial. Multiple anchors were intended to serve as alternative illustrations for that level of effectiveness and did not all have to be met. Further, raters were instructed to focus on the actual actions the participants took rather than the outcome.

If both raters indicated that they were unable to rate a given response because of insufficient information, that participant was removed from the dataset. If one rater believed the response was insufficient but the other provided a rating, a third rater independently reviewed the response and either rated the response or removed the participant. Three participants were removed because both raters indicated insufficient information, and three additional ones were removed when the independent rater verified that there was insufficient information, forming a final sample size of 199.

Last, if the ratings made by two corresponding raters were more than two points apart, the third rater again rated the response, and that rating replaced the more discrepant one. Mean interrater reliability was .71 for BD1-1 and .72 for BD1-2 (single rater), and .83 and .84 respectively after applying the Spearman-Brown formula (Spearman, 1904) to adjust for combining two sets of ratings. Complete details and results of the decision rules are available from the first author.

Response Transformations

To identify each participant's effectiveness range, we subtracted BDI-3 (worst handling of a difficult customer) from BDI-2 (best handling). We set a minimum difference of 1.0 for inclusion in the analysis, mainly because the resulting quartiles would have been too small with differences smaller than 1. A total of 70 participants were dropped because their range was too small, leaving a sample size of 129. Follow-up investigation revealed the underlying cause. Although these participants really did "blow it" as worded in BDI-3, they followed up and were able to salvage the situation. Our raters were influenced by these follow-up actions and increased their BDI-3 ratings as a result, thereby reducing the range below the 1.0 minimum.

For participants whose range was at least 1.0, we broke that range down into quartiles for each participant and then determined in which quartile their BDI-1 response fell. Quartiles 2 and 3 were classified as typical whereas Quartile

5 Although counterbalancing the order of study stimuli is a common psychometric practice, we always asked BDI-1 first to prevent contamination effects from BDI-2. For example, if we had asked BDI-2 first, priming would occur and that could easily have carried over into BDI-1. The order of the priming (BDI-2) and outright failure (BDI-3) questions could have been alternated.

4 was classified as maximal. Responses were transformed accordingly to 1 (typical) and 2 (maximal). This classification framework was based directly on our rating scale, the scale points for which functionally created four quadrants. Specifically, mean ratings could fall from 1.0 to 1.99, 2.0 to 2.99, 3.0 to 3.99, or 4.0 to 5.0. Scale points 2 and 3 reflected appropriate actions (to varying degrees) but not overly extensive effort. As noted earlier, 4 involved a "concerted" effort and 5 extraordinary effort. We noted further that this scheme resulted in the typical region being twice as large as the maximal region (i.e., two quartiles vs. one quartile).

Given the lack of guidance in the intraindividual performance literature regarding where the maximal region begins (as noted earlier), the use of quartiles seemed to balance (to a reasonable degree) the need to be somewhat stringent and yet have a sufficient number of maximal performance data points to analyze. We fully acknowledge that the choice to use quartiles was subjective, and encourage exploration of alternative classification schemes in future research including more stringent ones (e.g., top decile).

To illustrate the quartile process, one of our study participants had mean ratings of 2.3 for BDI-1, 3.3 for BDI-2, and 1.0 for BDI-3. Subtracting BDI-3 from BDI-2 resulted in an experiential range of 2.3 (i.e., 3.3-1.0). Dividing that range by four (i.e., 2.3 / 4) resulted in a quartile width of 0.575, and the resulting quartiles were 1.000[6] to 1.575 (1.000+.575) for Quartile 1, 1.576 to 2.150 (1.576+.575) for Quartile 2, 2.151 to 2.725 (2.151+.575) for Quartile 3, and 2.726 to 3.300 (2.726+.575) for Quartile 4. The mean rating for the primary question BDI-1 (2.3) fell into the third quartile, and thus it was classified as being a typical experience for this participant.

6 The lower end of the first quartile started with the BDI-3 rating of 1.0, and then the quartile width (.575) was added to it to compute the higher end of the first quartile. The second quartile started .001 higher than the end of the first quartile (its lower end), and then .575 was added to compute its higher end. Same for Q3 and Q4.

RESULTS

All results are displayed in Table 1. The first hypothesis was that the proportion of maximal responding for our primary question worded in the standard manner (BDI-1) would be significantly lower than the .481 value found by Huffcutt et al. (2020). There were 38 maximal and 54 typical responses, resulting in a maximal proportion of .413. This value is lower than their value but not as low as we had expected. A *z*-test of two independent proportions only just met the significance threshold ($z = 1.64$, $p = .05$). Although this hypothesis is statistically supported, we feel that the reported significance level only provides partial support given the modest magnitude of the difference.

The second hypothesis was that mean ratings for our primed question (BDI-2) would be significantly higher than mean ratings for the main study question (BDI-1). The mean rating for BDI-2 was 3.3 with a standard deviation of 0.79, whereas the mean rating for BDI-1 was 2.9 with a standard deviation of 0.77. A test of two dependent samples (one-tailed) was highly significant ($t = 5.42$, $p < .001$), providing strong and encouraging support for the hypothesis.

The third hypothesis was that the proportion of responses reaching a mean rating of 4.0 or higher would

TABLE 1.

Study Results by Hypothesis

Hypothesis 1: Comparison of maximal proportions (frequencies in parentheses)			
	Maximal	Typical	Significance
Huffcutt et al. (2020)	.481 (50)	51.9 (54)	
Current study	.413 (38)	58.7 (54)	$z = 1.64$, $p = .05$[a]
Hypothesis 2: Comparison of mean ratings (mean /standard deviation, N = 199)			
	BDI-2	BDI-1	Significance
Current study	3.3 / .79	2.9 / .77	$t = 5.42$, $p < .001$
Hypothesis 3: Proportion of mean ratings reaching 4.0 or higher (N = 199)			
	BDI-2	BDI-1	Significance
Current study	.258	.122	$z = 3.35$, $p < .05$
Hypotheses 4&5: Correlations with response length (N = 199)			
	BDI-2	BDI-1	
Mean ratings (MR)	.36, $p < .001$	.20, $p < .01$	

Note. [a]This significance test is for the difference in maximal proportions between Huffcutt et al. and the current study.

be significantly higher for the primed question (BDI-2) relative to the main study question (BDI-1). For BDI-2, 51 of the 199 participants (.258) had a mean rating of 4.0 or higher, whereas 24 of 199 (.122) did so for BDI-1. The difference between these two proportions (.258 vs. .122) was significant and in the prediction direction ($z = 3.35$, $p < .05$). It would appear that priming the wording of a question to enhance maximal responding does have a tendency to push more candidates into that top zone on the rating scale (i.e., 4.0 to 5.0).

The fourth hypothesis was that response length (word count) would correlate positively and significantly with mean interviewer ratings for the standard question (BDI-1). The correlation between mean BDI-1 ratings and word length was .20 (197 *df*, $p < .01$), which is significant and provides support for this hypothesis.

The fifth hypothesis was that response length would correlate positively and significantly with mean interviewer ratings for the primed question (BDI-2). The correlation between mean BDI-2 ratings and word length was .36 (197 *df*, $p < .001$), which is highly significant and provides strong support for this hypothesis. It would seem that word length is even more influential when priming is already in place.

DISCUSSION

A recent meta-analysis found a fully corrected mean criterion-related validity of .42 for structured interview formats, the highest among all predictors analyzed (Sackett et al., 2022). Although strong, a high proportion of performance variance remains unaccounted for by these interviews. We advocate for an expanded and stronger focus on new approaches and new directions for BDIs to capture more of this variance, particularly those that could enhance the consistency of maximal reporting. The long-term goal is to have every candidate provide an experience that portrays their highest (top-end) level of capability regardless of how well it compares to that of other candidates. (Alternatively, if assessment of typical performance is better matched with business and selection strategy, the goal would be to have every candidate describe an experience that reflects the average level of effort they tend to display on a day-to-day basis.)

We found empirical evidence supporting the efficacy of two factors for increasing maximal responding, priming and response length, which is very promising. Priming (rewording questions to focus directly on top-end response) is particularly promising given the common convention that BDI questions focus mainly on the problem (e.g., difficult customer) rather than the nature of the approach taken (absolute best handling) and that it is an easy change to implement. Encouraging longer responses appears to compliment priming. We now review results for each hypothesis in this study individually.

Hypothesis 1: Proportion of Maximal Responding With BDI-1

Results of this investigation provide a second estimate of the proportion of BDI responses that reflect maximal capability, one based on more standardized methodology and a very diverse working adult participant pool. Although our estimate was lower (.413 vs .481), both estimates are still in the .4-.5 range and suggest a somewhat robust population tendency. Such a level could prove problematic, as mixing typical and maximal responding makes it more difficult to conduct accurate selection and identify the truly top candidates. Finding ways to drive BDI responding consistently toward maximal performance should be one of the very top priorities for future research.

Such resilient mixing may not be entirely surprising, however. Limitations in memory recall noted earlier (e.g., four-slot working memory, experiences stored as scattered fragments) suggest that the human memory system really does not possess the innate "mental horsepower"' to respond effectively to BDI questions, at least not consistently across candidates. These limitations are exacerbated by the implicit expectation that candidates begin their responses shortly after a question is read, as failure to do so can lead to negative attributions by the interviewer (Brosy et al., 2020). It is entirely possible that achieving more consistent responding will have to involve a reduction in cognitive load.

Hypothesis 2: Higher Means Ratings for the Primed Question (BDI-2 vs. BDI-1)

It is surprising that the issue of rewording BDI questions to enhance maximal (or typical) responding has not yet been raised strongly in the BDI literature. An implicit norm seems to have emerged to word questions in a manner that focuses on the difficult problem or situation but not on finding one where the candidate displayed top-end capability to resolve it. We urge interview developers to consider shifting the paradigm to give full attention to both aspects of performance.

Unfortunately, priming BDI questions by itself does not appear to go far enough to achieve consistent maximal responding. Raising the mean rating (in comparison to BDI-1) from 2.9 to 3.3 is clearly a positive outcome and in the right direction, but the magnitude (0.4 increase) seems somewhat modest. Priming might operate to reduce cognitive load at least somewhat, as it narrows the long-term memory search parameters.

Hypothesis 3: The Proportion of BDI-2 Responses Reaching 4.0 or Higher

The proportion of mean ratings in the primed question reaching 4.0 or higher was more than double that of the standard question. Although having roughly a quarter of

mean ratings reach 4.0 or higher is probably acceptable, it is important to emphasize that this level is likely to continue to rise as maximal responding becomes progressively more consistent (e.g., as cognitive loading is reduced) and may reach a level where variability becomes too constrained. We encourage exploration of alternate rating scales that would allow the top end to be expanded. For instance, scales utilizing more than five points (e.g., seven, nine) could be considered or, alternately, some type of point system.

Hypotheses 4 and 5: Response Length (i.e., Word Count)

Although the benefits of longer responses appear promising, particularly for the primed question, it is important to note that we did not experimentally manipulate length. Rather, we used length as it naturally occurred, and as such, there is the potential for a covariate to account for the influence of word length. Participants who are more conscientious, for example, might give longer responses and be more effective in handling difficult customers.

We encourage research on response length at several levels. One stream could investigate whether there are covariates influencing both length and responding, and if so, how they might be used to enhance maximal responding across candidates. Another stream could attempt an experimental treatment of response length, including both the original response and the response provided after follow up (probing).

Study Limitations

As always, limitations should be noted. First, our interviews were administered completely online. Although asynchronous formats are becoming increasingly popular (Griswold et al., 2022), the degree to which our results generalize to face-to-face or virtual interviews is uncertain. Further, we did not impose a time limit on responding online, and the resulting range of completion time was quite large. We encourage exploration of time management in future research, including the effects of imposing a time limit.

Second, our participants did not have the same motivation to perform well as would real job candidates. Our sense was that use of impression management (IM) tactics was limited in our dataset, but such tactics (see Bourdage et al., 2020; Melchers et al., 2020) could easily emerge more strongly if actual job candidates were tested. Increased use of IM could increase the prevalence of maximal responding. Further, it could cause overestimation of top-end points in individual effectiveness ranges if candidate embellishment implies capabilities exceeding their true levels.[7]

Third, our methodology for classifying participant responses as typical or maximal was reasonable but could certainly be refined. We classified participant responses as maximal if they fell within the top quartile (25%) of the effectiveness range for that person. If we had used a more stringent classification point (e.g., top decile), the prevalence of maximal responding might easily have gone down even lower than 41.3%, painting an even bleaker picture of maximal responding in BDIs. In retrospect, analyzing multiple classifications simultaneously (e.g., top quartile and top decile) would have been optimal, but we leave that for future research.

Fourth, and relatedly, a number of participants did not reach the 1.0 minimum for the difference between BDI-2 and BDI-3, and thus could not be analyzed for typical versus maximal responding. We encountered an unexpected issue in that a number of participants did indeed fail to deal effectively with their difficult customer upfront (sometimes spectacularly) but then followed up and were able to salvage the situation (at least to some degree). Such an occurrence is common, as some difficult situations can be salvaged whereas others cannot. In hindsight, we should have worded BDI-3 to ask for a time they totally failed *and* were not able to recover.

Finally, we based computation of each participant's effectiveness range on only two experiences, one targeting high-end capability and one targeting low-end capability. Utilizing multiple questions to determine each endpoint should result in more stable estimates, much like adding additional items on a personality measure increases its reliability (Spearman, 1904). For instance, participants could be asked to describe two experiences where they handled a difficult customer in an exceptional way and two where they really failed (and couldn't recover).

Notwithstanding these limitations, we believe this study is unique in several ways and adds incrementally to our understanding of the BDI technique. In addition to confirming that typical and maximal responding are indeed mixed somewhat randomly, we explored and found empirical support for two mechanisms that appear to help induce more maximal responding (i.e., priming and response length). To our knowledge, we are the first in the BDI literature to raise the issue of rewording questions to prime maximal responding. Much work yet remains to be done, however, and our hope is that this study sparks new lines of research on this most remarkable structured interview format.

7 We wish to thank an anonymous reviewer for pointing out this important and realistic possibility.

REFERENCES

Aarts, H., Custers, R., & Holland, R. W. (2007). The nonconscious cessation of goal pursuit: When goals and negative affect are coactivated. Journal of Personality and Social Psychology, 92(2), 165–178.

Aguinis, H., & O'Boyle, E., Jr. (2014). Star performers in twenty-first century organizations. Personnel Psychology, 67(2), 313-330. http://dx.doi.org/10.1111/peps.12054

Bangerter, A., Corvalan, P., & Cavin, C. (2014). Storytelling in the selection interview? How applicants respond to past behavior questions. Journal of Business and Psychology, 29(4), 593-604. http://dx.doi.org/10.1007/s10869-014-9350-0

Barns, C. M., & Morgeson, F. P. (2007). Typical performance, maximal performance, and performance variability: Expanding our understanding of how organizations value performance. Human Performance, 20(3), 259-274. https://psycnet.apa.org/doi/10.1080/08959280701333289

Birt, J. (2022, May 18). How to use the STAR interview response technique. Indeed. https://www.indeed.com/career-advice/interviewing/how-to-use-the-star-interview-response-technique

Bourdage, J. S., Schmidt, J., Wiltshire, J., Nguyen, B., & Lee, K. (2020). Personality, interview performance, and the mediating role of impression management. Journal of Occupational and Organizational Psychology, 93(3), 556-577. https://bpspsychub.onlinelibrary.wiley.com/doi/10.1111/joop.12304

Bragger, J., Kutcher, E. J., Schettino, G., Muzyczyn, B., Farago, P., & Fritzky, E. (2016). Interview and cognitive performance: Does structure reduce reliance on selection batteries, and can explanation of purpose improve it? Performance Improvement Quarterly, 29(2), 97-124. https://doi.org/10.1002/piq.21218

Brosy, J., Bangerter, A., & Ribeiro, S. (2020). Encouraging the production of narrative responses to past-behaviour interview questions: Effects of probing and information. European Journal of Work and Organizational Psychology, 29(3), 330-343. http://dx.doi.org/10.1080/1359432X.2019.1704265

Cowan, N. (2010). The magical mystery four: How is working memory capacity limited and why? Current Directions in Psychological Science, 19(1), 51-57. http://dx.doi.org/10.1177/0963721409359277

Culbertson, S. S., Weyhrauch, W. S., & Huffcutt, A. I. (2017). A tale of two formats: Direct comparison of matching situational and behavior description interview questions. Human Resource Management Review, 27(1), 167-177. http://dx.doi.org/10.1016/j.hrmr.2016.09.009

Deadrick, D. L., & Gardner, D. G. (2008). Maximal and typical measures of job performance: An analysis of performance variability over time. Human Resource Management Review, 18(3), 133-145. https://psycnet.apa.org/doi/10.1016/j.hrmr.2008.07.008

Griswold, K. R., Phillips, J. M., Kim, M. D., Mondragon, N., Liff, J., & Gully, S. M. (2022). Global differences in applicant reactions to virtual interview synchronicity. International Journal of Human Resource Management, 33(15), 2991-3018. https://doi.org/10.1080/09585192.2021.1917641

Huffcutt, A. I., & Howes, S. S. (2023). Episodic recall from long-term memory: Theoretical overview and implications for behavior description interviews. Manuscript submitted for publication.

Huffcutt, A. I., Howes, S. S., Dustin, S. L., Chmielewski, A. N., Marshall, C. A., Metzger, R. L., & Gioia, V. P. (2020). Empirical assessment of typical versus maximal responding in behavior description interviews. Human Performance, 33(5), 447-467. http://dx.doi.org/10.1080/08959285.2020.1812075

Huffcutt, A. I., & Roth, (1998). Racial group differences in employment interview evaluations. Journal of Applied Psychology, 83(2), 179-189. https://psycnet.apa.org/doi/10.1037/0021-9010.83.2.179

Janz, T. (1982). Initial comparison of patterned behavior description interviews versus unstructured interviews. Journal of Applied Psychology, 67(5), 577-580. https://psycnet.apa.org/doi/10.1037/0021-9010.67.5.577

Judiesch, M. K., Schmidt, F. L., & Mount, M. K. (1992). Estimates of the dollar value of employee output in utility analyses: An empirical test of two theories. Journal of Applied Psychology, 77(3), 234-250. http://dx.doi.org/10.1037/0021-9010.77.3.234

Klehe, U.-C., & Latham, G. (2006). What would you do—really or ideally? Constructs underlying the behavior description interview and the situational interview in predicting typical versus maximum performance. Human Performance, 19(4), 357-382. http://dx.doi.org/10.1207/s15327043hup1904_3

Latham, G. P., Saari, L. M., Pursell, E. D., & Campion, M. A. (1980). The situational interview. Journal of Applied Psychology, 65(4), 422-427. https://doi.org/10.1037/0021-9010.65.4.422

Lievens, F., Lang, J. W. B., De Fruyt, F., Corstjens, J., Van de Vijver, M., & Bledow, R. (2018). The predictive power of people's intraindividual variability across situations: Implementing whole trait theory in assessment. Journal of Applied Psychology, 103(7), 753-771. http://dx.doi.org/10.1037/apl0000280

Loftus, E. F. (1995). Memory malleability: Constructivist and fuzzy-trace explanations. Learning and Individual Differences, 7(2), 133-137. http://dx.doi.org/10.1016/1041-6080(95)90026-8

Marcus, B., Goffin, R. D., Johnston, N. G., & Rothstein, M. (2007). Personality and cognitive ability as predictors of typical and maximum managerial performance. Human Performance, 20(3), 275-285. https://psycnet.apa.org/doi/10.1080/08959280701333362

Melchers, K. G., Roulin, N., & Buehl, A.-K. (2020). A review of applicant faking in selection interviews. International Journal of Selection and Assessment, 28(2), 123-124. http://dx.doi.org/10.1111/ijsa.12280

Morgeson, F. P., Reider, M. H., & Campion, M. A. (2005). Selecting individuals in team settings: The importance of social skills, personality characteristics, and teamwork knowledge. Personnel Psychology, 58(3), 583-611. http://dx.doi.org/10.1111/j.1744-6570.2005.655.x

O'Boyle, E., Jr., & Aguinis, H. (2012). The best and the rest: Revisiting the norm of normality of individual performance. Personnel Psychology, 65(1), 79-119. http://dx.doi.org/10.1111/j.1744-6570.2011.01239.x

Roulin, N., & Krings, R. (2016). When winning is everything: The relationship between competitive worldviews and job applicant faking. Applied Psychology: An International Review, 65(4), 643-670. http://dx.doi.org/10.1111/apps.12072

Sackett, P. R. (2007). Revisiting the origins of the typical—maximal performance distinction. Human Performance, 20(3), 179-185. http://dx.doi.org/10.1080/08959280701332968

Sackett, P. R., Zhang, C., Berry, C. M., & Lievens, F. (2022). Revisiting meta-analytic estimates of validity in personal selection: Addressing systematic overcorrection for restriction of range. Journal of Applied Psychology, 107(11), 2040-2068. http://dx.doi.org/10.1037/apl0000994

Sackett, P. R., Zedeck, S., & Fogli, L. (1988). Relations between measures of typical and maximum job performance. Journal of Applied Psychology, 73(3), 482-486. http://dx.doi.org/10.1037/0021-9010.73.3.482

Schmidt, F. L., & Hunter, J. E. (1983). Individual differences in productivity: An empirical test of estimates derived from studies of selection procedure utility. Journal of Applied Psychology, 68(3), 407-414. http://dx.doi.org/10.1037/0021-9010.68.3.407

Spearman, C. (1904). The proof and measurement of association between two things. American Journal of Psychology, 15, 72-101. https://psycnet.apa.org/doi/10.2307/1412159