Computerized methods for collecting confidence ratings: Task influences on patterns of responding

K. Andrew DeSoto

Retrospective confidence ratings and other judgments frequently are collected in computer-based psychology studies, but little research has investigated whether the method with which these ratings are collected influences the resulting data. To explore whether different confidence rating entry methods elicit different responses, 96 subjects were tested in a recognition memory paradigm. To rate confidence in recognition decisions from 0 - 100, half of the subjects used the numeric keypad on the keyboard to respond whereas the other half used an on-screen slider. Notably, whereas subjects using the numeric keypad frequently chose to enter confidence ratings divisible by 5 and 10, subjects using the slider showed no such preference but instead were more likely to accept the slider default value (i.e., 50) for each trial. The method with which confidence ratings are collected may have unintended consequences on confidence rating data and their interpretation.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1432v1 | CC-BY 4.0 Open Access | rec: 14 Oct 2015, publ: 14 Oct 2015

Computerized methods for collecting confidence ratings: Task influences in patterns of responding

K. Andrew DeSoto[1]

[1] Department of Psychological & Brain Sciences, Washington University in St. Louis, St. Louis, MO, USA. K. Andrew DeSoto is now at the Association for Psychological Science, Washington, DC, USA.

Corresponding Author:
K. Andrew DeSoto [1]
1133 15th Street NW, Suite 1000, Washington, DC, 20005, USA
Email address: adesoto@psychologicalscience.org
Web: http://www.andydesoto.com/
Twitter: @kadesoto

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1432v1 | CC-BY 4.0 Open Access | rec: 14 Oct 2015, publ: 14 Oct 2015

Abstract

Retrospective confidence ratings and other judgments frequently are collected in computer-based psychology studies, but little research has investigated whether the method with which these ratings are collected influences the resulting data. To explore whether different confidence rating entry methods elicit different responses, 96 subjects were tested in a recognition memory paradigm. To rate confidence in recognition decisions from 0 - 100, half of the subjects used the numeric keypad on the keyboard to respond whereas the other half used an on-screen slider. Notably, whereas subjects using the numeric keypad frequently chose to enter confidence ratings divisible by 5 and 10, subjects using the slider showed no such preference but instead were more likely to accept the slider default value (i.e., 50) for each trial. The method with which confidence ratings are collected may have unintended consequences on confidence rating data and their interpretation.

Computerized Methods for Collecting Confidence Ratings:

Task Influences on Patterns of Responding

Psychologists commonly present stimuli and collect data using computerized methods. As a result, simple studies traditionally conducted using pencil and paper can now be designed in a variety of ways and with a cornucopia of user interface components. In addition, these studies can be coded with any number of programming languages ranging from ActionScript (e.g., Weinstein, 2012) to Python (e.g., Peirce, 2007), all of which feature idiosyncrasies. As a result, two psychologists can implement even the simplest experimental task very differently.

For instance, in the area of metacognition (see Dunlosky & Metcalfe, 2008, for a review), subjects often make judgments and predictions that reflect the monitoring and control processes that occur during learning (Nelson & Narens, 1990). One judgment frequently used in the laboratory is the *retrospective confidence rating* (often referred to as *confidence*). In a typical experiment employing confidence ratings, a subject sits at a computer and is asked to learn a list of words. After a brief delay, the subject is given a recognition test in which he or she is asked if a displayed word had been studied earlier in the experiment. After the subject responds (e.g., with "old" or "new"), he or she is prompted: *How confident are you that your previous answer is correct?* The subject rates his or her confidence on a provided scale and proceeds with the rest of the test. Confidence ratings may be collected dozens of times over the course of a single laboratory experiment, and most researchers assume that these ratings are accurate representations of the subjective sense of confidence actually experienced (but see Roediger & DeSoto, 2015; Roediger, Wixted, & DeSoto, 2012, for a discussion).

There are many different ways an experimenter can collect confidence ratings in a computerized fashion, however. To name several variables, confidence ratings can be collected on a 1 - 7 scale or a 0 - 100 scale, paced by either the experimenter or the subject, and entered by the subject in several different ways (see DeSoto, 2014, for further discussion). Subjects can even make ratings jointly (Bahrami, Olsen, Latham, Roepstorff, Rees, & Frith, 2010). Regrettably, little scientific effort has been expended to compare these methods. This is well put by Vickers (2001, as cited in Koriat, 2012, p. 80), who noted, "Despite its practical importance and pervasiveness, the variable of confidence seems to have played a Cinderella role in cognitive psychology — relied on for its usefulness, but overlooked as an interesting variable in its own right." Vickers' concern seems appropriate: Recent research suggests the collection method for cognitive and metacognitive judgments affects the judgments that are made as well as the distribution and variance of those judgments (e.g., Benjamin, Tullis, & Lee, 2013; but see Kellen, Klauer, & Singmann, 2013).

For example, Ariel, Al-Harthy, Was, and Dunlosky (2011) presented subjects with three items of varying memorability (i.e., difficulty) arranged horizontally on a computer screen and asked them to choose the order in which they wanted each item to appear in a later study phase. Ariel and colleagues found that study choices were inconsistent with existing theories of self-regulated learning, which suggest that item difficulty should affect study choices (e.g., Metcalfe, 2009). Regardless of the difficulty of the items presented, however, English-speaking subjects overwhelmingly chose items in left-to-right order. In contrast, Arabic-speaking subjects, who read right-to-left, tended to select items to study in a right-to-left order. The implication was that subjects selected items in the same direction in which they read text, even though this was not likely to be the most effective way to optimize performance on the task (for further research, see Ariel & Dunlosky, 2013).

The work of Ariel and colleagues illustrates an important point: When subjects participate in an experimental task that requires many metacognitive judgments to be made, they engage in certain behaviors to reduce the cognitive demands (e.g., working memory load) required by the task (as suggested by Krosnick, 1991). Specifically, subjects often select the first response that seems reasonable, and may also have a tendency to accept the "status quo" response — that is, a default or middle-of-the-road response. Subjects "may give this answer without any retrieval or judgement, simply because it appears to be a reasonable answer" (p. 219).

Although initial evidence suggests differences in task influence cognitive and metacognitive performance on a general level, very few studies have investigated specifically how collection methods alter the resulting confidence rating data. Therefore, the purpose of this study was to investigate differences in how subjects' responses varied when they were asked to enter their confidence ratings in one of two ways. In this study, subjects participated in a standard recognition memory paradigm and provided retrospective confidence ratings. To investigate differences in responses resulting from the method of assessing confidence, subjects either entered their confidence rating using (1) the numeric keypad on their keyboard or (2) a graphical on-screen slider.

The prediction was that the two entry methods would result in different responses due to affordances or biases elicited by each task. Specifically, the hypothesis was that the numeric keypad entry method would encourage subjects to enter numbers ending in zero or five (e.g., much preferring an answer of 80 over 81), because mentally reducing the 100-point confidence scale to a 10 or 20-point scale would be less cognitively demanding. In contrast, it was predicted that a slider entry method would not show biases toward these numbers but would instead encourage choosing the slider's default value for each trial (e.g., 50 if the slider was set to 50 for each trial). With the slider entry method, accepting the default value is even less demanding than selecting a rating ending in zero or five.

Method

A total of 96 Washington University in St. Louis students participated for either course credit or payment. (This number was chosen in advance of data collection to be consistent with earlier experiments.) Subjects studied 150 words taken from category norms (Van Overschelde, Rawson, & Dunlosky, 2004) and then took an immediate recognition memory test over 300 items (150 targets and 150 lures). In this test, subjects were presented with one item at a time and asked to decide whether each item was old or new. After making a recognition decision, subjects were asked to rate their confidence that their prior decision was correct on a scale from 0 (*not at all confident*) to 100 (*entirely confident*).

Subjects were assigned to one of two experimental groups: the *numeric keypad entry* group or the *slider entry* group. (Data for these two groups were collected at different times during the academic year and previously published as Experiment 1 and Experiment 2 by Roediger and DeSoto, 2014; see also DeSoto & Roediger, 2014, for a similar design). Entry method was the only difference between the two groups. Subjects in the numeric keypad group were asked to type in their confidence rating (from 0 - 100) using the numeric keypad portion of the computer keyboard. After typing in a number, subjects pressed the Enter or Return key to submit their judgment and proceed to the next trial. On the other hand, subjects in the slider entry group were presented with a visual slider on the screen. Confidence ratings in increments of 10 were marked on the scale, which ran across approximately half of the length of the screen. Subjects were asked to move the slider, which began at a default position of 50, to the desired

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1432v1 | CC-BY 4.0 Open Access | rec: 14 Oct 2015, publ: 14 Oct 2015

point on the scale. Subjects then clicked a button on the screen to submit their judgment and proceed to the next trial. In both conditions, text appeared on the screen reminding subjects that a zero meant *not at all confident* and a 100 meant *complete confidence*. See Figure 1 for an illustration of what subjects in each of these two conditions saw for each trial.

There were no data exclusions or other experimental manipulations not reported in the study. Basic trial data and item-level data were collected (e.g., reaction time, word frequency) but were not analyzed because the focus was on confidence. The study was approved by the Washington University in St. Louis Institutional Ethics Board (#201112008) and conformed to Standard 8 of the American Psychological Association's Ethical Principles of Psychologists and Code of Conduct. The data for this project are public and available on *fig**share*** (DeSoto, 2015).

Results

Average confidence ratings were significantly higher in the keypad group (M = 71, SD = 11.84) than in the slider group (M = 65, SD = 11.80), $t(94) = 2.16$, $p = .03$, 95% CI [0.41, 10.00], $d = 0.10$. Although this effect size was small, it warranted further investigation of the distribution of confidence ratings as a function of experimental group, which is shown in Figure 2.

Visual inspection of these two distributions reveals three main observations. First, an extreme number of confidence ratings of 100 were provided by subjects in both conditions. This is a typical finding in research investigating rating scales (e.g., Mickes, Hwe, Wais, & Wixted, 2011). Second, and more interesting, is that subjects using the keypad entry method appeared to choose confidence ratings that were divisible by five or 10 more often than subjects in the slider entry group. This is confirmed by an independent-samples t-test comparing the percentage of responses assigned to a number divisible by 10 in the keypad entry group (M = .78, SD = .20) to the percentage of responses divisible by 10 in the slider entry group (M = .50, SD = .20), $t(94)$ =

6.95, $p < .001$, 95% CI [.20, .37], $d = 1.43$. Turning to all responses divisible by five, the difference is even more striking: Significantly more responses were assigned to a number divisible by five in the keypad entry group ($M = .92$, $SD = .15$) than in the slider entry group, ($M = .54$, $SD = .19$), $t(94) = 11.16$, $p < .001$, 95% CI [.32, .45], $d = 2.30$ (note that all responses divisible by 10 are also divisible by 5).

Third, Figure 2 also shows that confidence ratings of 50 occurred much more frequently in the slider entry group ($M = .20$, $SD = .22$) than in the keypad entry group ($M = .09$, $SD = .09$). This was also confirmed by an independent-samples t-test, $t(94) = 3.27$, $p = .002$, 95% CI [.04, .18], $d = 0.67$.

As a measure of the relative metamemory accuracy of subjects' confidence ratings, gamma correlations were computed for each subject (see Nelson, 1984). Gamma correlations, like Pearson correlations, range from -1.00 to 1.00 and represent the degree to which confidence is associated with accuracy within individuals, with higher correlations indicating a greater degree of correspondence. Subjects appeared to have equivalent relative metamemory accuracy in both the keypad entry group ($M = .38$, $SD = .19$) and the slider entry group ($M = .38$, $SD = .20$). Indeed, an independent-samples t-test failed to identify a significant difference, $t(94) = 0.06$, $p = .947$, 95% CI [-.08, .08], $d = 0.01$.

Several asides: Interestingly, throughout 14,000 judgments, no subject in the keypad entry group provided a confidence rating of 61 or 71, but all possible ratings (0 - 100) were made at least once by subjects in the slider entry group. Additionally, the keypad group was programmed originally to accept any user input, regardless of validity. Thus, some of the resulting data, amounting to 1.5% of the trials, were erroneous (e.g., some subjects entered "900" when they probably intended to enter "90") and had to be excluded. Those programming keypad entry methods in the future should take precautions to ensure that this does not occur.

Discussion

In sum, subjects who were asked to enter confidence with the numeric keypad overwhelmingly chose to enter confidence ratings divisible by five or 10 (doing so more than 92% of the time). On the other hand, subjects who were prompted to enter confidence with the slider showed a preference for the default value of the slider (i.e., reporting a confidence of 50 approximately 20% of the time). These results strongly imply that biases driven by collection method are at play when subjects report confidence in a recognition memory paradigm.

It is possible that when subjects repeatedly made 0 - 100 confidence ratings using the numeric keypad, it became difficult to keep track of (or perhaps even distinguish between) different levels of confidence provided (e.g., Keren, 1991). As a result, subjects may have mentally abridged the rating scale, choosing to respond only in multiples of five or 10. Subjects that responded with the slider, however, were provided with a default value. Because providing a confidence rating of 50 only entailed one action (i.e., clicking on the submit button) instead of two (i.e., moving the slider to the desired position and then clicking on the submit button), subjects seemed to differentially prefer ratings of 50. It is unclear whether this is due to an anchor-and-adjust heuristic (e.g., Tversky & Kahneman, 1992), preference for the status quo response, or mere apathy. Investigating responses as a function of overall experiment length could begin to identify the degree to which these factors are involved (see Hanczakowski, Zawadzka, Pasek, & Higham, 2013, for another look at scaling biases).

Future research will be necessary to determine the best methods to prevent task biases from influencing data (or even whether such biases need to be eliminated in the first place).

Recent studies conducted in our laboratory using a slider entry method (e.g., DeSoto & Roediger, 2014) have required subjects to move the slider at least once before the response can be submitted, preventing subjects from easily selecting the default confidence rating for each trial. Other alternatives may involve setting the slider default to 0, 100, or even a different value for each trial (although this method would also contribute additional error variance). A still better approach may involve the use of a *visual analog scale* (e.g., Marsh-Richard, Hatzis, Mathias, Venditti, & Dougherty, 2009; Reips & Funke, 2008). These scales typically require subjects to select a point on a line to indicate their judgment (i.e., rather than clicking and dragging a slider) and are likely to be less susceptible to biases in responding because there is no default value.

Acknowledgments

I thank Roddy Roediger, the Washington University Memory Lab, Joe Bernardi, Becky Koenig, and the attendees of the 43rd annual meeting of the Society for Computers in Psychology for their comments on this research.

References

Ariel R, Al-Harthy IS, Was CA, Dunlosky J. (2011). Habitual reading biases in the allocation of study time. Psychonomic Bulletin and Review 18: 1015-1021. DOI:10.3758/s13423-011-0128-3.

Ariel R, Dunlosky J. (2013). When do learners shift from habitual to agenda-based processes when selecting items for study? Memory & Cognition 41: 416-428. DOI:10.3758/s13421-012-0267-4.

Bahrami B, Olsen K, Latham PE, Roepstorff A, Rees G, Frith CD. (2010). Optimally interacting minds. Science 329: 1081-1085. DOI:10.1126/science.1185718.

Benjamin AS, Tullis JG, Lee JH. (2013). Criterion noise in ratings-based recognition: Evidence from the effects of response scale length on recognition accuracy. Journal of Experimental Psychology: Learning, Memory, and Cognition 39: 1601-1608. DOI:10.1037/a0031849.

DeSoto KA. (2014). Collecting confidence ratings in cognitive psychology experiments: Investigating the relationship between confidence and accuracy in memory. Brindle, P, editor. SAGE cases in methodology. Thousand Oaks: SAGE. DOI:10.4135/978144627305013507683.

DeSoto, KA. (2015). Computerized methods for collecting confidence ratings: Task influences on patterns of responding: Dataset. ***figshare***. DOI:10.6084/m9.figshare.1572182.

DeSoto KA, Roediger HL. (2014). Positive and negative correlations between confidence and accuracy for the same events in recognition of categorized lists. Psychological Science 25: 781-788. DOI:10.1177/0956797613516149.

Hanczakowski M, Zawadzka K, Pasek T, Higham P. (2013). Calibration of metacognitive judgments: Insights from the underconfidence-with-practice effect. Journal of Memory and Language 69: 429-444. DOI:10.1016/j.jml.2013.05.003.

Kellen D, Klauer KC, Singmann H. (2013). On the measurement of criterion noise in signal detection theory: Reply to Benjamin (2013). Psychological Review 120: 727-730. DOI:10.1037/a0033141.

Keren G. (1991). Calibration and probability judgments: Conceptual and methodological issues. Acta Psychologica 77: 217-273. DOI:10.1016/0001-6918(91)90036-Y.

Koriat A. (2012). The self-consistency model of subjective confidence. Psychological Review, 119: 80-113. DOI:10.1037/a0025648.

Krosnick JA. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. Applied Cognitive Psychology 5: 213-236. DOI:10.1002/acp.2350050305.

Marsh-Richard DM, Hatzis ES, Mathias CW, Venditti N, Dougherty DM. (2009). Adaptive visual analog scales (AVAS): A modifiable software program for the creation, administration, and scoring of visual analog scales. Behavior Research Methods 41: 99-106. DOI:10.3758/BRM.41.1.99.

Metcalfe J. (2009). Metacognitive judgments and control of study. Current Directions in Psychological Science 18: 159-163. DOI:10.1111/j.1467-8721.2009.01628.x.

Mickes L, Hwe V, Wais PE, Wixted, JT. (2011). Strong memories are hard to scale. Journal of Experimental Psychology: General 140: 239-257. DOI:10.1037/a0023007.

Nelson TO. (1984). A comparison of current measures of the accuracy of feeling-of-knowing predictions. Psychological Bulletin 95: 109-133. DOI:10.1037/0033-2909.95.1.109.

Peirce JW. (2007). PsychoPy — Psychophysics software in Python. Journal of Neuroscience Methods 162: 8-13. DOI:10.1016/j.jneumeth.2006.11.017.

Reips U-D, Funke F. (2008). Interval-level measurement with visual analogue scales in Internet-based research: VAS Generator. Behavior Research Methods 40: 699-704. DOI:10.3758/BRM.40.3.699.

Roediger HL, DeSoto, KA. (2014). Confidence in memory: Assessing positive and negative correlations. Memory 22: 76-91. DOI:10.1080/09658211.2013.795974

Roediger HL, DeSoto, KA. (2015). Understanding the relation between confidence and accuracy in reports from memory. In Lindsay D, Kelley C, Yonelinas A, Roediger H, editors. Remembering: Attributions, processes, and control in human memory. New York, NY: Psychology Press. pp. 347-367.

Roediger HL, Wixted JH, DeSoto, KA. (2012). The curious complexity between confidence and accuracy in reports from memory. In Nadel L, Sinnott-Armstrong W, editors. Memory and law. Oxford: Oxford University Press. pp. 84-118. DOI:10.1093/acprof:oso/9780199920754.001.0001

Tversky A, Kahneman D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. Journal of Risk and Uncertainty 5: 297-323. DOI:10.1007/BF00122574.

Weinstein Y. (2012). Flash programming for the social and behavioral sciences: A simple guide to sophisticated online surveys and experiments. Thousand Oaks: SAGE. DOI:10.4135/9781452244099.

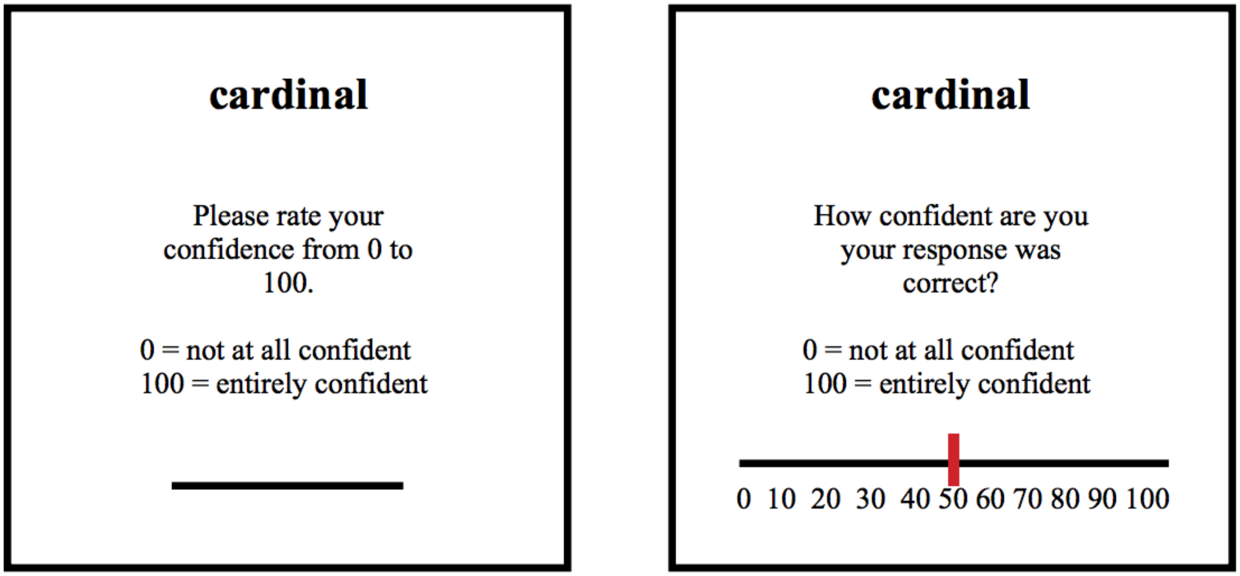
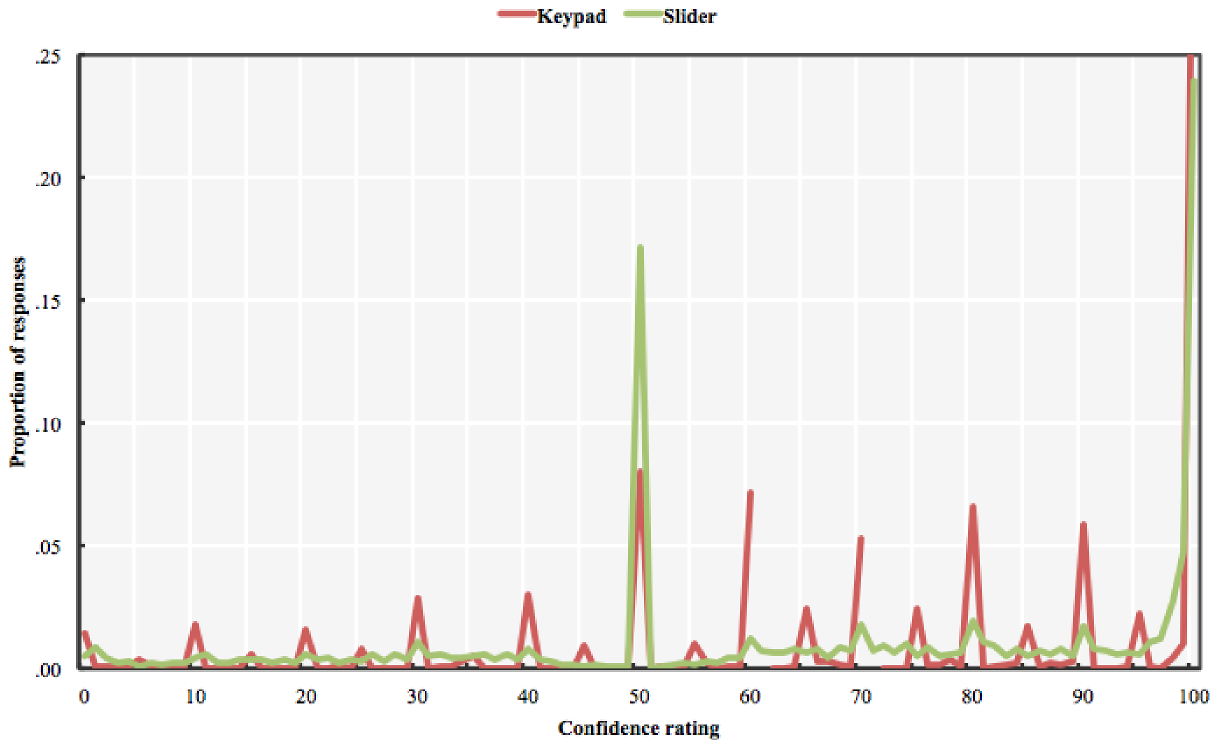


Figure 1. A depiction of the two types of confidence rating entry methods used in the numeric keypad group (left panel) and slider group (right panel).



274 *Figure 2*. The proportion of total confidence ratings assigned to each value in the 0 - 100 range
275 by subjects in the keypad entry group and subjects in the slider entry group. (No ratings of 61 or
276 71 were provided by subjects in the keypad entry group.)