2014 Human Factors *Prize Finalist*

Are Well-Calibrated Users Effective Users? Associations Between Calibration of Trust and Performance on an Automation-Aided Task

Stephanie M. Merritt, **Deborah Lee**, **Jennifer L. Unnerstall**, and **Kelli Huber**, University of Missouri, St. Louis

Objective: We present alternative operationalizations of trust calibration and examine their associations with predictors and outcomes.

Background: It is thought that trust calibration (correspondence between aid reliability and user trust in the aid) is a key to effective human–automation performance. We propose that calibration can be operationalized in three ways. Perceptual accuracy is the extent to which the user perceives the aid's reliability accurately at one point in time. Perceptual sensitivity and trust sensitivity reflect user adjustment of perceived reliability and trust as the aid's actual reliability changes over time.

Method: One hundred fifty-five students completed an X-ray screening task with an automated screener. Awareness of the aid's accuracy trajectory and error type was examined as predictors, and task performance and aid failure detection were examined as outcomes.

Results: Awareness of accuracy trajectory was significantly associated with all three operationalizations of calibration, but awareness of error type was not when considered in conjunction with accuracy trajectory. Contrary to expectations, only perceptual accuracy was significantly associated with task performance and failure detection, and combined, the three operationalizations accounted for only 9% and 4% of the variance in these outcomes, respectively.

Conclusion: Our results suggest that the potential importance of trust calibration warrants further examination. Moderators may exist.

Application: Users who were better able to perform the task unaided were better able to identify and correct aid failure, suggesting that user task training and expertise may benefit human–automation performance.

Keywords: trust, automation, awareness, performance, error

Address correspondence to Stephanie M. Merritt, University of Missouri–St. Louis, 421 Stadler Hall, St. Louis, MO 63121, USA; e-mail: merritts@umsl.edu.

HUMAN FACTORS
Vol. 57, No. 1, February 2015, pp. 34–47
DOI: 10.1177/0018720814561675

INTRODUCTION

As automation becomes more widespread, appropriate reliance on automated systems becomes increasingly important for safety and effectiveness. Inappropriate reliance on automation can take one of two forms. *Misuse* refers to overreliance on automation, whereas *disuse* refers to a tendency to underrely on automation, even when increased reliance would improve performance (Parasuraman & Riley, 1997). In order to achieve maximum safety and performance, both misuse and disuse should be avoided, and *appropriate reliance* should be achieved (e.g., Dzindolet, Peterson, Pomranky, Pierce, & Beck, 2003).

To achieve appropriate reliance, it has been suggested that operators must correctly "calibrate" their level of trust in the system. *Calibration of trust* has been defined as "the correspondence between the person's trust in the automation and the automation's capabilities" (Lee & See, 2004, p. 55), and facilitating calibration has been suggested as a goal for system designers (Lee & See, 2004; Parasuraman & Miller, 2004; Wiegmann, 2002). The general notion is that when trust corresponds to the aid's reliability, users will rely on the automation more appropriately, and human–automation performance will be maximized.

The overall aim of this study is twofold. First, we theoretically explore the construct of calibration and operationalize it in three ways. Second, we examine predictors and outcomes of trust calibration. A model of our hypotheses is presented in Figure 1.

Calibration of Trust

As previously described, *calibration* refers to the correspondence between an aid's performance capabilities and the user's level of trust in the aid. However, difficulties are inherent

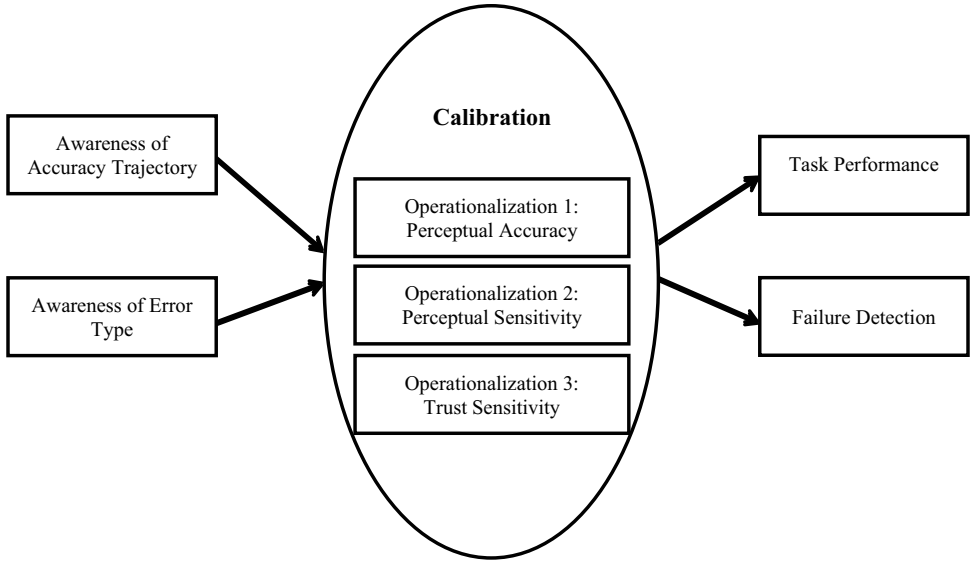

Figure 1. Hypothesized relationships.

in the operationalization of calibration. Trust may operate on a different latent scale than automation reliability, making it difficult to directly compare the two. Trust is an attitudinal construct reflecting the degree to which a user will rely on the automation to achieve his or her goals under conditions of uncertainty (Lee & See, 2004). In contrast, perceived reliability is often conceptualized as a percentage accuracy rate (e.g., the automation is 99% reliable). This difference makes a simple comparison of scores on trust and perceived reliability scales inappropriate. These scales are unlikely to achieve measurement invariance (e.g., Vandenberg & Lance, 2000), and thus, mean scores on those scales cannot be compared.

We propose that calibration may be operationalized in three ways. The extent to which each of the three operationalizations is associated with expected predictors and outcomes of calibration is assessed. The first operationalization, *perceptual accuracy*, is defined as the extent to which user perceptions of the aid's reliability reflect the aid's actual level of reliability. This operationalization is consistent with the definition of calibration proposed by Parasuraman and Miller (2004), who suggested that users should possess accurate beliefs about the automation's reliability, and it has been used in some past research (e.g., McGuirl & Sarter, 2006). To operationalize perceptual accuracy, the user's perceptions of the aid's reliability are compared with the aid's actual reliability. Greater correspondence between actual and perceived reliability at a given time point reflects greater calibration, whereas greater discrepancies between actual and perceived reliability reflect lesser calibration. Perceptual accuracy can thus be measured by assessing a user's beliefs about aid reliability and comparing those beliefs to the aid's actual reliability level. This assessment could be accomplished using a difference score or, when used as a predictor, more complex techniques, such as polynomial regression (e.g., Edwards & Parry, 1993).

The second operationalization of calibration incorporates the notion of changes in reliability. When decisions are made under uncertainty, automation reliability may change as conditions fluctuate. Thus, calibration may be operationalized as the degree to which users' perceptions adjust as aid reliability changes. We define our second operationalization, *perceptual sensitivity*, as the extent to which the user's perceptions of automation reliability change in correspondence with changes in the aid's actual reliability. Thus, whereas perceptual accuracy can be measured at a single point in time, perceptual sensitivity reflects within-person changes over time. Any measure of perceptual sensitivity should reflect

within-person associations between changes in aid reliability and changes in perceptions of aid reliability. Thus, in order to measure perceptual sensitivity, a repeated-measures design is required in which the correspondence between actual and perceived reliability is assessed at varying levels of actual reliability. In past research, this assessment has been done at the group level by comparing the mean level of perceived aid reliability across conditions and before versus after a change in reliability (Wiegmann, Rich, & Zhang, 2001). However, we focus on calibration at the individual level, such that we assess the degree to which a given user is perceptually sensitive. To assess individual-level perceptual sensitivity, one option is to employ multilevel modeling whereby the Level 1 equation reflects the within-person correspondence of actual reliability with perceived reliability. Perfect calibration would be indicated by a finding of $\beta = 1.00$, suggesting a 1:1 correspondence between changes in reliability and perceived reliability. Another option would be to compute the within-person correlation of perceived and actual aid reliability, whereby correlations closer to 1.00 indicate greater perceptual sensitivity.

The third and final proposed calibration operationalization, *trust sensitivity*, reflects the extent to which a user's *trust* changes as the automation's actual reliability level changes. Thus, similar to perceptual sensitivity, the assessment of trust sensitivity requires a repeated-measures design in which actual reliability varies. Although it has been measured at the group level in previous research (Wiegmann et al., 2001), trust sensitivity can also be operationalized at the individual level as the within-person correspondence between changes in trust and changes in reliability. Like the previous operationalization, it can be assessed using multilevel modeling, whereby a finding of $\beta = 1.00$ would suggest that a one-unit change in actual reliability was associated with a one-unit change in user trust. It can also be assessed by computing the within-person correlation of actual reliability and trust.

Of the three operationalizations, trust sensitivity is most consistent with the definition of calibration as the correspondence between automation reliability and trust. However, the disadvantage is that because of the differences in the latent scale between reliability and trust, one cannot determine the degree to which trust is "correctly" calibrated. Individuals with a great deal of trust sensitivity (meaning that their trust changes dramatically as actual reliability changes) might actually either overreact or underreact to the changes, thereby resulting in suboptimal automation reliance decisions. In contrast, the advantage of the perceptual accuracy and perceptual sensitivity operationalizations is that they allow the researcher to assess the degree to which beliefs accurately reflect reality.

Predictors of Trust Calibration

Some research has suggested that providing information regarding the aid's reliability level or aid confidence may increase calibration (Lee & See, 2004; Sheridan & Parasuraman, 2006; St. John, Smallman, Manes, Feher, & Morrison, 2005). The notion is that users who have more accurate beliefs about aid performance characteristics should have better calibration of trust. Therefore, we investigate the relationships of two aspects of user awareness (awareness of the aid's accuracy trajectory and awareness of the type of errors made) as potential predictors of each operationalization of calibration.

Awareness of accuracy trajectory. Individuals may vary in the extent to which they are consciously aware of an automated aid's performance characteristics, including whether the aid's reliability is increasing or decreasing over time. We expect that individuals who are consciously aware of the way in which the aid's reliability is changing will be better able to determine when the aid is erring and therefore better able to calibrate trust. Thus, we hypothesize the following:

> *Hypothesis 1*: Awareness of the aid's accuracy trajectory will be positively associated with calibration.

Awareness of error type. Two error types, false alarms and misses, may have differential effects on user behavior. In signal detection terms, false-alarm errors are those in which the automation indicates that the signal is present

when it is not, and misses are errors in which the automation indicates that the signal is absent when it is actually present. Users may vary in the extent to which they recognize that the aid is making false-alarm versus miss errors. To the extent that individuals correctly recognize the type of error the aid makes, we expect that they will also be able to better determine when the aid is erring and better able to calibrate trust. Therefore, we hypothesize the following:

Hypothesis 2: Awareness of error type will be positively associated with calibration.

Outcomes of Calibration

To our knowledge, researchers have not yet empirically examined the effect size of the relationships between trust calibration and outcomes. In order to do so, we select two relevant outcomes: detection of aid failures and overall task performance.

Failure detection. Ideally, users rely on automation when it is correct but reject it when it is incorrect. Individuals tend to adopt one of two strategies: probability matching or maximization (e.g., Wiegmann, 2002). When probability matching, users match their reliance rates to their perceptions of aid reliability such that if they believe the aid is correct 90% of the time, they rely on it 90% of the time. Using the maximization strategy with a 90% reliable aid, individuals always rely on the aid because this reliance is likely to maximize the number of correct decisions made. Probability-matching strategies may be more common (Walker, 1996); however, neither strategy necessarily results in avoidance of rare errors. For probability matching to be effective, users must rely on the automation for the correct 90% of trials; otherwise their accuracy rate will be under 90%. Thus, it is important that users can identify aid errors in order to correct them, while relying on the automation the rest of the time. To the extent that users have calibrated levels of trust, they may also be able to identify aid errors. We therefore hypothesize the following:

Hypothesis 3: Trust calibration will be significantly associated with failure detection.

Task performance. Finally, we examined overall task performance. Whereas correct disagreements on error trials may be the key outcome in safety-critical tasks, in noncritical tasks, overall task performance may be of greater concern. Although task performance is expected to be significantly associated with user ability to detect failures, it will be considered a separate outcome. We hypothesize the following:

Hypothesis 4: Trust calibration will be significantly associated with task performance.

METHOD

A 2 (aid accuracy trajectory: increasing or decreasing) × 2 (aid error type: false alarm or miss) between-subjects manipulation was used. Participants were 156 college students at an urban university in the midwestern United States and were recruited through course participation in the psychology subject pool or through contact with course instructors. Most received extra credit for participation. One participant completed the study multiple times, and we were unable to accurately match that participant's survey and performance data, so the participant was excluded. The final sample thus consisted of 155 participants. Demographic information was obtained for 154 of these; this sample was 71% female. Racially, the sample was 56.5% White, 21.4% Black, 11.0% Asian American, and 0.6% multiracial, and 10.5% declined to report race. Participants were provided with a definition and examples of automation (e.g., GPS, ATM) and reported how much experience they had using automation on a 4-point scale. The mean score was 2.99, with 74.2% of participants reporting that they either had "moderate" (3) or "a great deal" of (4) experience with automation.

Procedure

The study was conducted online. Participants completed self-report measures assessing demographics and individual differences not relevant to the present hypotheses. Next, they completed a four-block X-ray screening task similar to those used by Merritt and Ilgen (2008) and Merritt (2011). Due to the relative simplicity of the

task, no formal training was provided. Further, participants were given no information about aid reliability prior to this task; therefore their calibration was not affected by any prior information. Participants were simply instructed to visually search for guns and knives in the luggage images and that an automated aid would provide advice, which they could either choose to accept or reject when making a final decision. Each block contained a series of 20 X-ray slides of luggage. For each image, participants first provided their initial opinion ("search" or "clear"), then received the advice of a fictitious automated screener. The aid recommended that the participant either search (i.e., the aid believes a weapon is present) or clear (i.e., it believes no weapon is present) each bag. After viewing the aid's advice, the participant made a final decision and received feedback on whether a weapon was in fact present in the image. Following each 20-slide block, participants rated their current levels of trust in the aid as well as its perceived reliability. A more detailed task description, including slide order, is available from the corresponding author. When the entire task was complete, awareness of aid accuracy trajectory and error type was assessed. No time limits were imposed; most participants completed the study within 1 hr.

Manipulations

Accuracy trajectory. Participants were randomly assigned to the increasing or decreasing reliability condition. In the increasing condition, the aid was correct 80%, 85%, 90%, and 95% of the time, respectively. In the decreasing condition, the aid's reliability decreased from 95% to 80% over the course of the blocks. In both conditions, the X-ray images viewed were identical and were presented in the same order.

Error type manipulation. Participants were randomly assigned to experience only one type of error. This design provides us with greater experimental control and is consistent with past research (e.g., Dixon & Wickens, 2006; Dixon, Wickens, & McCarley, 2007; Rice, 2009). In the false-alarm condition, the aid's errors suggested that a weapon was present when, in fact, the bag was safe. In the miss condition, a weapon was present but the aid suggested that the participant clear the bag.

Measures

Trust. Trust was assessed using a three-item version of the scale used by Merritt, Heimbaugh, LaChapell, and Lee (2013). Items included "I believe the automatic screener is a competent performer," "I trust the automatic screener," and "I can depend on the automatic screener." Items were on a 5-point Likert-type scale ranging from *strongly disagree* to *strongly agree*. This scale was administered after each of the four task blocks. The internal consistency reliabilities for this scale were $\alpha = .93$ in Blocks 1 through 3 and $\alpha = .96$ in Block 4.

Perceptual accuracy. We calculated perceptual accuracy as the mean difference between perceived and actual reliability for each individual across the four task blocks. We then took the absolute value of that score. Thus, perceptual accuracy scores are exclusively positive, with greater scores representing lower accuracy.

Note that the perceived reliability scale that participants saw used response options in 5% increments ranging from *less than 50%* to *100%* reliable, which we translated into numbers (e.g., 80%) in order to calculate accuracy. Thus, although mean values reflect perceived compared to actual percentages, individuals' specific discrepancies were always in 5% increments.

Perceptual sensitivity. This variable assessed the degree to which the user's perceptions of aid reliability changed as actual reliability changed. Because it involves repeated observations nested within individuals, this variable was operationalized in a multilevel model that accounts for such nesting. For tests of Hypotheses 1 and 2, perceptual sensitivity was operationalized as the Level 1 association of perceived and actual reliability in a multilevel equation, reflected by the individual's personalized beta weight between actual reliability and perceived reliability. However, when testing Hypotheses 3 and 4, we needed to calculate a single value for this variable in order to use it as a predictor of task performance and failure detection. Thus, for Hypotheses 3 and 4, the variable was operationalized as the within-person correlation of actual and perceived reliability.

Trust sensitivity. This variable assessed the degree to which the user's trust changed as aid reliability changed. This variable was operationalized in a manner identical to perceptual

sensitivity with the exception that the variable of interest was trust instead of perceived reliability.

Awareness of aid accuracy trajectory and error type. Awareness was assessed with two multiple-choice items consisting of four options each. The first asked the participant to describe how the aid's performance accuracy was changing over time, and four options were presented (increasing, decreasing, staying the same, or changing unpredictably). The second asked participants to identify the type of error the aid made when it made a mistake, and four options were presented (missed a weapon that was there, gave a false alarm when no weapon was there, some of both, or neither—the aid did not make any mistakes). Each item was coded according to the participant's condition such that correct responses were coded 1 (aware) and incorrect responses were coded 0 (unaware).

Failure detection. Participants' ability to detect aid failures was calculated as the percentage of time users' final decisions disagreed with faulty automation advice (correct disagreements) divided by the percentage of incorrect disagreements. Final decisions were used for this calculation because users received the aid's advice after the initial decision was made; thus, only the final decision could be considered a rejection of the aid's advice.

The ratio of correct disagreements to incorrect disagreements was used so that indiscriminate disuse of the aid would not be classified as failure detection. For example, a participant who disagreed with the automation on 80% of the trials when it gave incorrect advice and only 20% of the trials when it gave correct advice would have a score of 0.8/0.2 = 4. In contrast, a participant who disagreed with the aid on 50% of the trials when it gave incorrect advice and on 50% of the trails when it gave correct advice would have a score of 0.5/0.5 = 1. Thus, higher values reflect a greater ability to detect aid failures. Scores below 1 indicated that participants disagreed with the automation more often when it was correct than when it was incorrect.

Task performance. Participants accumulated points for each correct final decision and were penalized for each incorrect final decision. Therefore, task performance scores reflect the joint performance of the human–automation team. Ten points were awarded for a hit, and 10 were deducted for a miss. Five points were awarded for a correct rejection, and five points were deducted for a false positive. Higher point awards and penalties were used when weapons were present to reflect the importance of preventing weapons from entering an aircraft. Overall task performance was measured as the total number of points accumulated across the four task blocks. Performance scores were examined for outliers (>3 standard deviations from the mean). Two participants performed extremely poorly (>5 standard deviations below the mean) and were removed from the analyses involving task performance.

RESULTS

Descriptive Statistics

Table 1 displays the means and standard deviations for each scale. In Table 2, average discrepancies of perceived and actual reliability are displayed by block and condition. As shown in Table 2, participants tended to underestimate the aid's reliability, which is consistent with some past research (e.g., Wiegmann, 2002; Wiegmann & Cristina, 2000). Significant differences were found between the increasing and decreasing accuracy trajectory conditions in Blocks 2 through 4. Underestimation of aid performance was significantly greater in the increasing condition than in the decreasing condition. For error type, false alarms were associated with significantly more accurate perceptions of reliability overall and in Blocks 1 and 4 compared to misses.

Table 3 displays the results of the manipulation checks by condition. Table 4 displays the bootstrapped correlations among the three operationalizations of trust calibration. For these calculations, perceptual sensitivity and trust sensitivity were calculated as the within-person correlation of actual reliability with perceived reliability and trust, respectively. As shown, the correlations among the three operationalizations were relatively low, suggesting that these three variables seem to be distinct.

Table 5 provides descriptive information regarding the benefits of the aid. Initial (unassisted) decision correctness is compared with final (assisted) decision correctness. On average, receiving the aid's advice increased the percentage

TABLE 1: Descriptive Statistics Overall and/or by Task Block, as Applicable

	Overall			Block 1			Block 2			Block 3			Block 4		
Variable	*M*	*SD*	*n*	*M*	*SD*	*n*	*M*	*SD*	*n*	*M*	*SD*	*n*	*M*	*SD*	*n*
Trust				3.28	.97	151	3.22	.96	148	3.23	.93	146	3.33	1.00	146
Perceived reliability				83.75	15.55	148	84.16	13.06	148	84.04	12.60	146	84.24	12.63	145
Perceptual accuracy[a]	9.30	8.65	153	9.56	9.93	148	8.18	9.76	148	9.38	10.00	146	10.41	10.61	145
Ac aware	0.44	0.50	147												
Er aware	0.58	0.50	144												
Failure detection	3.27	2.88	142												
Performance	266.29	53.86	143												

Note. Ac aware = awareness of accuracy trajectory; Er aware = awareness of error type; both are scored 0 = unaware, 1 = aware.
[a]Perceptual accuracy is a mean based on the absolute value of each participant's discrepancy between perceived and actual aid reliability such that all scores are positive and greater values reflect greater inaccuracies in either direction.

TABLE 2: Discrepancies Between Perceived and Actual Aid Reliability by Block and Condition

	Average		Block 1		Block 2		Block 3		Block 4	
Condition	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Overall	–3.55	11.46	–3.75	13.29	–3.24	12.32	–3.56	13.26	–3.52	14.47
Increasing	–7.02	12.98	–4.32	15.29	–5.58	14.32	–8.03	14.29	–10.33	14.05
Decreasing	0.15	8.16	–3.18	10.99	–0.70	9.15	1.29	10.10	3.79	10.98
False alarms	–1.41	10.01	–1.09	11.97	–1.74	10.94	–2.32	12.34	0.00	12.40
Misses	–5.46	12.36	–6.08	14.00	–4.56	13.35	–4.68	14.08	–6.62	15.51

Note. Means between the increasing and decreasing accuracy conditions were significant ($p < .05$) overall and for Blocks 2 through 4. Means were significantly different between error type conditions ($p < .05$) overall and at Blocks 1 and 4. Negative means represent underestimates of aid reliability and positive means represent overestimates of aid reliability.

of correct decisions by approximately 5% regardless of condition.

Effects of Awareness

Hypotheses 1 and 2 suggested that awareness of accuracy trajectory and awareness of error type, respectively, would be significantly associated with calibration. We begin by operationalizing calibration as perceptual accuracy. A linear regression was performed in which the two awareness variables were entered simultaneously. Awareness of accuracy trajectory associated with perceptual accuracy to a degree that approached significance ($\beta = .18$, $p = .06$), providing some evidence consistent with Hypothesis 1. However, the effect of awareness of error type ($\beta = .03$, $p = .79$) was not significant, failing to support Hypothesis 2 for perceptual accuracy.

Next, perceptual sensitivity was assessed by examining the within-person association of actual reliability with perceived reliability, with greater correspondence between perceived and actual reliability indicating greater perceptual sensitivity. Because within-person associations (e.g., perceived and actual reliability across the four blocks) were nested within individuals, this relationship was assessed using the Level 1 (within-person) equation in multilevel modeling.

TABLE 3: Accuracy Trajectory and Error Type Awareness: Percentages of Responses by Condition

	Decreasing Accuracy		Increasing Accuracy	
Response	False Alarms	Misses	False Alarms	Misses
Accuracy awareness ("Throughout the study, the ABI's accuracy . . .")				
Increased	6.0	2.7	33.3[a]	42.5[a]
Decreased	45.5[a]	54.1[a]	8.3	2.5
Stayed the same	39.4	32.4	41.7	42.5
Varied unpredictably	9.1	10.8	16.7	12.5
Error awareness ("When the ABI made a mistake, it . . .")				
Missed a weapon that was there	6.2	70.3[a]	11.4	60.0[a]
Gave a false alarm when no weapon was there	50.0[a]	2.7	51.4[a]	2.5
Some of both	43.8	27.0	34.3	37.5
Neither—the ABI did not make any mistakes	0.0	0.0	2.9	0.0

Note. Percentages reflect valid responses (exclude missing data).
[a]Indicates the correct response by condition.

TABLE 4: Correlations Among Perceptual Accuracy, Perceptual Sensitivity, and Trust Sensitivity

Variable	Perceptual Accuracy	Perceptual Sensitivity	Trust Sensitivity
Perceptual accuracy[a]	—		
Perceptual sensitivity [95% CI]	–.28* [–.47, –.10]	—	
Trust sensitivity [95% CI]	–.09 [–.29, .09]	.33* [.13, .51]	—

Note. Perceptual accuracy is a discrepancy score such that higher scores reflect lower accuracy. Correlations were bootstrapped with 1,000 random samples. Standard errors for all correlations were $SE = .10$, $N = 112$.
*$p < .05$

These multilevel analyses were performed on the portion of the sample that lacked missing awareness data ($N = 137$). Although null models are not presented here, they were examined prior to conducting the multilevel analyses. Within persons, perceived and actual aid reliability were significantly associated ($\pi_{10} = .55$, $p < .01$), suggesting that participants perceived the changes in aid reliability, consistent with past work (e.g., Wiegmann et al., 2001). The intraclass correlation (ICC) estimate indicated that 71% of the variance in perceived reliability was between persons.

Hypotheses 1 and 2 were tested by adding the between-person variables of awareness of accuracy trajectory and awareness of error type at Level 2 of the multilevel equation and examining the interaction coefficients. The hypothesis was that participants who were correctly aware of the aid's characteristics would have greater correspondence between actual aid reliability and perceived aid reliability. Both awareness variables were coded such that 0 = not aware and 1 = aware. Chi-square difference tests indicated that a random error term for the interactions fit the data best, $\Delta\chi^2(2) = 20.22$. Therefore, the random interaction error term was used.

Assessment of the parameter estimates in Table 6 indicates that accuracy trajectory and error type awareness had significant main effects on user perceptions of aid reliability, such that those who correctly identified these factors had significantly lower estimates of aid reliability. Also, a significant interaction was found between actual aid reliability and awareness of accuracy trajectory ($\beta_{11} = .66$, $p < .01$), supporting Hypothesis 1 for perceptual sensitivity. Participants who correctly identified whether the

TABLE 5: Means and Standard Deviations for Percentage Correct Decisions Unassisted and Assisted by Condition

Condition	Unassisted (Initial Decision)	Assisted (Final Decision)
Increasing + false alarms	.76 (.08)	.81 (.07)
Increasing + misses	.75 (.07)	.80 (.06)
Decreasing + false alarms	.74 (.10)	.81 (.05)
Decreasing + misses	.77 (.08)	.83 (.06)

TABLE 6: Associations of Awareness With Perceptual Sensitivity

Fixed Effect	Parameter	Coefficient	Standard Error	*p*
For intercept (main effects of awareness on perceived reliability)				
Intercept	β_{00}	7.16	.29	<.01
Ac aware	β_{01}	–1.07	.46	.02
Er aware	β_{02}	–1.28	.45	.01
For slope (effects of awareness on perceptual sensitivity)				
Intercept	β_{10}	0.15	.13	.25
Ac aware	β_{11}	0.66	.12	<.01
Er aware	β_{12}	0.17	.13	.20

Note. Ac aware = awareness of accuracy trajectory; Er aware = awareness of error type. Perceptual sensitivity is the association of aid reliability and trust, so the cross-level interaction effects of awareness represent associations of awareness with perceptual sensitivity.

aid's accuracy was increasing or decreasing had greater perceptual sensitivity than those who did not. No significant interaction was found for awareness of error type, failing to support Hypothesis 2 for perceptual sensitivity. However, this may be due to multicollinearity with awareness of accuracy trajectory; when awareness of error type was tested in isolation, its effect was significant ($\beta = .32$, $p < .01$). This finding suggests that awareness of error type may be significantly linked to perceptual sensitivity when tested in isolation, but its effects may be eclipsed by those of awareness of accuracy trajectory.

Finally, we tested Hypotheses 1 and 2 for trust sensitivity. Trust sensitivity was assessed by examining the Level 1 (within-person) relationship of actual aid reliability and trust. Within persons, the mean association of actual reliability and trust was $\pi_{10} = .15$, $p < .01$, indicating a relatively small but significant association of aid reliability and trust. ICC indicated that 68% of the variance was between persons, and a random interaction term was indicated, $\Delta\chi^2(2) = 24.80$.

As shown in Table 7, both awareness variables had significant direct effects on trust, such that awareness was associated with lower trust. In addition, awareness of accuracy trajectory had a significant association with trust sensitivity, such that sensitivity was significantly greater for those who correctly identified the aid's performance trajectory ($\beta_{11} = .18$, $p < .01$). Hypothesis 1 was supported for trust sensitivity. No significant effect was found for awareness of error type, failing to support Hypothesis 2 for trust sensitivity. However, as with perceptual sensitivity, the effect of error type awareness was significant when tested in isolation, without awareness of accuracy trajectory ($\beta = .12$, $p = .03$).

Outcomes: Task Performance and Failure Detection

We next addressed Hypotheses 3 and 4, the associations of calibration with failure detection

TABLE 7: Associations of Awareness With Trust Sensitivity

Fixed Effect	Parameter Label	Coefficient	Standard Error	*p*
For intercept (direct effects of awareness on trust)				
Intercept	β_{00}	3.48	.13	<.01
Ac aware	β_{01}	–0.49	.17	.01
Er aware	β_{02}	–0.42	.17	.02
For slope (effects of awareness on trust sensitivity)				
Intercept	β_{10}	0.03	.04	.52
Ac aware	β_{11}	0.18	.05	<.01
Er aware	β_{12}	0.08	.05	.12

Note. Ac aware = awareness of accuracy trajectory; Er aware = awareness of error type. Trust sensitivity is the association of aid reliability and trust, so the cross-level interaction effects of awareness represent associations of awareness with trust sensitivity.

TABLE 8: Correlations Between Trust Calibration Components and Outcomes: Task Performance and Failure Detection

	Task Performance		Failure Detection	
Variable	*r*	95% CI	*r*	95% CI Lower
Perceptual accuracy	–.32[a]	[–.47, –.13]	–.19[a]	[–.28, –.08]
Perceptual sensitivity	–.00	[–.25, .23]	–.29	[–.51, .08]
Trust sensitivity	–.05	[–.25, .15]	–.11	[–.31, .08]

[a]95% confidence interval excluded zero. Perceptual accuracy is a discrepancy measure; thus, lower scores on this measure indicate more accurate perceptions.

and task performance. Consistent with the tests of Hypotheses 1 and 2, perceptual accuracy was operationalized as the absolute value of the discrepancy between perceived and actual aid reliability, averaged across the task blocks. However, for these analyses, we needed to create single scores for perceptual sensitivity and trust sensitivity. Perceptual sensitivity was therefore operationalized as the within-person correlation between perceived and actual reliability across the four task blocks (M = .53, SD = .53, N = 130). Trust sensitivity was operationalized via the within-person correlation of trust and actual reliability across the four task blocks (M = .36, SD = .63, N = 127). For these analyses, missingness was treated using pairwise deletion.

First, we examined correlations between each of the three operationalizations of calibration and the two outcomes, which were bootstrapped using 1,000 simple samples. As shown in Table 8, the confidence intervals for perceptual accuracy excluded zero, such that greater perceptual accuracy (coded in terms of lower absolute value discrepancy between perceptions and reality) was associated with significantly higher task performance and failure detection. However, no significant relationships were found for perceptual sensitivity or trust sensitivity. Thus, Hypotheses 3 and 4 were supported for perceptual accuracy only.

In order to assess the three operationalizations' relative contributions to the outcomes of interest, linear regressions were performed in which the three operationalizations were entered simultaneously. For task performance (Table 9), calibration accounted for approximately 9% of the variance (p < .01), and only perceptual accuracy

TABLE 9: Regression of Task Performance on Calibration Operationalizations

	Unstandardized Coefficients		Standardized Coefficients		
Variable	*B*	*SE*	Beta	*t*	*p*
(Constant)	291.91	10.59		27.56	<.01
Perceptual accuracy[a]	–2.74	0.76	–.35	–3.59	<.01
Perceptual sensitivity	–7.09	10.20	–.07	–0.70	.49
Trust sensitivity	–4.26	8.70	–.05	–0.49	.63

[a]Perceptual accuracy is a discrepancy measure; thus, lower scores on this measure indicate more accurate perceptions.

TABLE 10: Regression of Failure Detection on Calibration Operationalizations

	Unstandardized Coefficients		Standardized Coefficients		
Variable	*B*	*SE*	Beta	*t*	*p*
(Constant)	3.99	.63		6.32	<.01
Perceptual accuracy[a]	–0.08	.05	–.18	–1.81	.07
Perceptual sensitivity	0.10	.61	.02	0.17	.87
Trust sensitivity	–0.26	.52	–.05	–0.49	.63

[a]Perceptual accuracy is a discrepancy measure; thus, lower scores on this measure indicate more accurate perceptions.

achieved statistical significance ($p < .01$). For failure detection (Table 10), the model accounted for only 4% of the variance ($p = .23$), and none of the three calibration operationalizations achieved significance, although perceptual accuracy was marginally significant. These results suggest that of the three operationalizations of calibration, perceptual accuracy seemed to have the strongest associations with outcomes, although these associations were weaker than expected.

Supplemental Analyses

We conducted some exploratory analyses. First, we examined the direct association of the two awareness variables with task performance. Neither awareness of accuracy trajectory ($t = –1.12$, $p = .26$) nor awareness of error type ($t = –1.44$, $p = .15$) was significantly associated with performance. Awareness of accuracy trajectory was marginally associated with failure detection ($t = 1.71$, $p = .09$), such that those who correctly identified the accuracy trajectory ($M = 3.74$, $SD = 3.66$) trended toward better failure identification than those who did not ($M = 2.92$, $SD = 2.05$). However, awareness of error type was not significantly associated with failure detection ($t = –1.60$, $p = .11$).

The strongest predictor of failure detection in our data set did not concern appraisals of the automation. Instead, it concerned participants' ability to perform the task unaided. Our measure of participants' unaided task ability, the degree to which they were initially correct (before receiving the automation's advice), was significantly associated with failure detection ($r = .47$, $p < .01$). Thus, it seems that participants who were better able to perform the task unassisted were also better able to identify aid failures.

DISCUSSION

It is widely assumed that accurate calibration of trust is key to improving the performance of human–automation teams. It seems intuitive that users who more correctly perceive the aid's ability should have higher task performance and greater ability to identify aid failures than those who do not. However, our results sug-

gest that the benefits of trust calibration, at least in some situations, could be less potent than previously believed. Past research has seldom assessed trust calibration, and to our knowledge, this is the first attempt to empirically examine relationship between individual-level calibration and performance outcomes. We presented three operationalizations of calibration based on theoretical discussions of the construct and found that although perceptual accuracy was moderately correlated with task performance and failure detection, these correlations were lower than expected. Neither perceptual sensitivity nor trust sensitivity was significantly correlated with either outcome. Furthermore, when combined, the three operationalizations accounted for only 9% of the variance in task performance and 4% of the variance in failure detection. These results are novel in the literature and seem to indicate that more attention to empirical testing of the calibration construct is warranted.

One interesting finding is that the three operationalizations of trust calibration demonstrated relatively low intercorrelations, suggesting that they seem to be measuring different phenomena. We believe that further attention to the construct of calibration and its measurement is warranted. Greater construct clarification is needed on issues such as whether within-person adjustments are essential to calibration or whether point-in-time assessments, such as perceptual accuracy, can also represent calibration. If a comparison of trust and aid reliability is essential to the definition of calibration, then what measurement techniques can be developed to compare these variables given their lack of measurement equivalence? How important to the construct is the ability to assess whether trust is "correctly" calibrated in a normative sense? In the present study, we presented three potential operationalizations of the construct that highlight potential disagreements on the theoretical nature of calibration. Future work should be undertaken with the goal of developing a unified definition of the calibration construct.

We next attempt to theorize why calibration may not have shown stronger associations with task performance and ability to identify aid errors in this study. One potential explanation may be that users employed a maximization strategy in which they always relied on the aid (Wiegmann et al., 2001). However, our data do not seem to support this hypothesis. In cases in which the user's initial opinion differed from the aid's, users switched their opinion to match the aid's only 45% of the time on average, suggesting that any tendency to use maximization strategies was not widespread in this sample.

A second potential explanation may be that although users are able to identify a general or overall trend in reliability across a sample of observations (particularly when given performance feedback), the ability to determine the accuracy of any individual decision is more error prone. Third, the null associations between trust sensitivity and outcomes could relate to the "latent scale" problem described in the Introduction. Although users may adjust their trust levels as reliability changes, they may not adjust the "correct" amount. Because trust and reliability are on different scales, it was not possible in this study to determine whether trust sensitivity was normatively correct; we were able to say only that some users were more or less sensitive to changes in reliability than others.

Limitations and Suggestions for Future Research

The present study is limited in that we examine trust calibration on a single task in a single setting with a single measure of trust. A search for moderators of the associations between awareness, calibration, and performance would be beneficial. Task characteristics, level of automation, participant expertise, perceived risk, and participant motivation could moderate the associations discovered here.

Further, evidence is accumulating to suggest that human interactions with automation may not often be driven by purely rational processes. A sense of competition (Beck, Dzindolet, & Pierce, 2007), emotions (Merritt, 2011), and implicit attitudes (Merritt et al., 2013) are among the less-rational factors that seem to influence human–automation interactions. Future research should continue to address the conditions under which user decision making may be driven by more or less rational processes in order to determine situations in which calibration is most effective.

In our sample, rates of awareness of accuracy trajectory (44%) and error type (58%) were

lower than expected and had no significant correlations with task performance and only a marginal association between awareness of accuracy trajectory and failure detection. More specific data may be necessary regarding the mental decision-making processes of individuals performing decision tasks that are aided by automation, such as processes related to information discounting. Qualitative research may be helpful in this regard. Such detailed data may allow us to discover why greater awareness of aid performance characteristics would not prove significantly beneficial.

Practical Applications

At least in some situations, calibration of trust may not be as closely tied to performance as previously thought. User task ability, however, demonstrated a moderate-to-strong association with ability to identify aid failures. Thus, organizations might consider the importance of user task training. Even when users perform in a supervisory capacity, increased task skill may reduce errors. Also, because false alarms were associated with greater perceptual accuracy, systems might be designed to err on the side of false alarms rather than misses, especially in cases when misses are particularly damaging (however, a cry-wolf effect could occur in which users begin to disregard system warnings; Wickens et al., 2009).

CONCLUSION

Much attention has been given to the notion of trust calibration in the literature. However, the findings of the current study suggest that calibration might have smaller effects on performance than previously believed. More attention to questions of when, why, and how calibration may vary in its effects would be useful in improving the safety and effectiveness of human–automated teams.

ACKNOWLEDGMENTS

This research was funded by a grant from the Air Force Research Lab. We thank Kent Coffel for assistance with database management and our colleagues at the University of Missouri–St. Louis for their input on earlier drafts of this project.

KEY POINTS

- Trust calibration can be operationalized as perceptual accuracy, perceptual sensitivity, or trust sensitivity.
- Awareness of whether the aid's accuracy was increasing or decreasing over the task was significantly associated with each of the three operationalizations of calibration.
- Only perceptual accuracy was significantly associated with task performance and ability to identify aid failures, and combined, the three operationalizations of calibration accounted for only 9% of the variance in task performance and 4% of the variance in failure detection.
- The strongest associate of failure detection was the user's ability to perform the task unassisted.

REFERENCES

Beck, H. P., Dzindolet, M. T., & Pierce, L. G. (2007). Automation usage decisions: Controlling intent and appraisal errors in a target detection task. *Human Factors*, *49*, 429–437. doi:10.1518/001872007x200076

Dixon, S. R., & Wickens, C. D. (2006). Automation reliability in unmanned aerial vehicle control: A reliance-compliance model of automation dependence in high workload. *Human Factors*, *48*, 474–486.

Dixon, S. R., Wickens, C. D., & McCarley, J. S. (2007). On the independence of compliance and reliance: Are automation false alarms worse than misses? *Human Factors*, *49*, 564–572.

Dzindolet, M. T., Peterson, S. A., Pomranky, R. A., Pierce, L. G., & Beck, H. P. (2003). The role of trust in automation reliance. *International Journal of Human–Computer Studies*, *58*, 697–718. doi:10.1016/S1071-5819(03)00038-7

Edwards, J. R., & Parry, M. E. (1993). On the use of polynomial regression equations as an alternative to difference scores in organizational research. *Academy of Management Journal*, *36*, 1577–1613.

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, *46*, 50–80.

McGuirl, J. M., & Sarter, N. B. (2006). Supporting trust calibration and the effective use of decision aids by presenting dynamic system confidence information. *Human Factors*, *48*, 656–665.

Merritt, S. M. (2011). Affective processes in human-automation interactions. *Human Factors*, *53*, 356–370.

Merritt, S. M., Heimbaugh, H., LaChapell, J., & Lee, D. (2013). I trust it, but I don't know why: Effects of implicit attitude toward automation on trust in an automated system. *Human Factors*, *55*, 520–534. doi:10.1177/0018720812465081

Merritt, S. M., & Ilgen, D. R. (2008). Not all trust is created equal: Dispositional and history-based trust in human-automation interactions. *Human Factors*, *50*, 194–210.

Parasuraman, R., & Miller, C. A. (2004). Trust and etiquette in high-criticality automated systems. *Communications of the ACM*, *47*, 51–55.

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors*, *39*, 230–253.

Rice, S. (2009). Examining single- and multiple-process theories of trust in automation. *Journal of General Psychology*, *136*, 303–319.

Sheridan, T., & Parasuraman, R. (2006). Human–automation interaction. In R. S. Nickerson (Ed.), *Reviews of human factors and ergonomics* (Vol. 1, pp. 89–129). Santa Monica, CA: Human Factors and Ergonomics Society.

St. John, M., Smallman, H. S., Manes, D. I., Feher, B. A., & Morrison, J. G. (2005). Heuristic automation for decluttering tactical displays. *Human Factors*, *47*, 509–525.

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, *3*, 4–70. doi:10.1177/109442810031002

Walker, J. T. (1996). *The psychology of learning*. Upper Saddle River, NJ: Prentice Hall.

Wickens, C. D., Rice, S., Keller, D., Hutchins, S., Hughes, J., & Clayton, K. (2009). False alerts in air traffic control conflict alerting system: Is there a "cry wolf" effect? *Human Factors*, *51*, 446–462.

Wiegmann, D. A. (2002). Agreeing with automated diagnostic aids: A study of users' concurrence strategies. *Human Factors*, *44*, 44–50.

Wiegmann, D. A., & Cristina, F. J., Jr. (2000). Effects of feedback lag variability on the choice of an automated diagnostic aid: A preliminary predictive model. *Theoretical Issues in Ergonomic Science*, *1*, 139–156.

Wiegmann, D. A., Rich, A., & Zhang, H. (2001). Automated diagnostic aids: The effects of aid reliability on users' trust and reliance. *Theoretical Issues in Ergonomics Science*, *2*, 352–367. doi:10.1080/14639220110110306

Stephanie M. Merritt is an associate professor of psychology at the University of Missouri–St. Louis. She received a PhD from Michigan State University in industrial/organizational psychology in 2007 and was the recipient of the Jerome H. Ely Human Factors Award for Volume 50.

Deborah Lee is a doctoral candidate at the University of Missouri–St. Louis in the industrial/organizational psychology program. She earned an MA in industrial/organizational psychology from Western Kentucky University in 2009 and a master's degree in industrial/organizational psychology from the University of Missouri–St. Louis in 2012.

Jennifer L. Unnerstall is a doctoral candidate in the industrial/organizational psychology program at the University of Missouri–St. Louis. She received a BA in psychology from Truman State University in 2009 and a master's degree in industrial/organizational psychology from the University of Missouri–St. Louis in 2012.

Kelli Huber is a graduate student at the University of Missouri–St. Louis pursuing a degree in industrial/organizational psychology. She received a BA in psychology from Saint Louis University in 2011 and a master's degree in industrial/organizational psychology from the University of Missouri–St. Louis in 2013.

Date received: October 30, 2013
Date accepted: October 31, 2014