technical note technical note

Scale Size of the Air Traffic Workload Input Technique (ATWIT): A Review of Research

Thomas Fincannon, PhD, Applied Research Associates, Inc.
Vicki Ahlstrom, FAA Human Factors Branch

November 2014

DOT/FAA/TC-TN-14/45

U.S. Department of Transportation
Federal Aviation Administration

William J. Hughes Technical Center
Atlantic City International Airport, NJ 08405

NOTICE

Technical Report Documentation Page

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
DOT/FAA/TC-TN-14/45		

4. Title and Subtitle	5. Report Date
Scale Size of the Air Traffic Workload Input Technique (ATWIT): A Review of Research	November 2014
	6. Performing Organization Code ANG-E25

7. Author(s)	8. Performing Organization Report No.
Thomas Fincannon, PhD, Applied Research Associates, Inc. Vicki Ahlstrom, FAA Human Factors Branch	DOT/FAA/TC-TN-14/45

9. Performing Organization Name and Address	10. Work Unit No. (TRAIS)
Federal Aviation Administration Human Factors Branch William J. Hughes Technical Center Atlantic City International Airport, NJ 08405	11. Contract or Grant No.

12. Sponsoring Agency Name and Address	13. Type of Report and Period Covered
Federal Aviation Administration William J. Hughes Technical Center Atlantic City Airport, NJ 08405	Technical Report
	14. Sponsoring Agency Code ANG-E2

15. Supplementary Notes

16. Abstract

Objective: This paper uses a review of previous studies to provide a recommendation for the optimal scale size of the Air Traffic Workload Input Technique (ATWIT). **Background:** The ATWIT is a measure of workload that was originally a 10-point scale, but subsequent research includes a 7-point variation of this scale. Scale size is known to impact assessment reliability, and more reliable scales produce stronger effect sizes and reduce costs that are associated with experimentation. Therefore, it is important to know whether the 7-point or 10-point version of the scale is more reliable. **Method:** The authors conducted a preliminary meta-analysis of 15 studies. The analysis examined correlations between ratings using the ATWIT and aircraft count (an objective measure of difficulty) to compare effect sizes across studies with a 7-point scale and a 10-point scale. **Results:** Findings indicated that the strength of the correlation between ATWIT ratings and aircraft count was greater for the 10-point version of the ATWIT than for the 7-point version. **Conclusion:** The 10-point scale appears to be more appropriate for the ATWIT than the 7-point scale. The authors recommend that researchers use a 10-point for the ATWIT, unless they have clear justification for deviating from this convention. The authors recommend further research to examine and control for the effects of potential confounds.

17. Key Words	18. Distribution Statement
Air Traffic Workload Input Technique Aircraft Count Meta-Analysis Workload Assessment Keypad	This document is available to the public through the National Technical Information Service, Alexandria, Virginia, 22312. A copy is retained for reference at the William J. Hughes Technical Center Library.

19. Security Classification (of this report)	20. Security Classification (of this page)	21. No. of Pages	22. Price
Unclassified	Unclassified	19	

THIS PAGE IS BLANK INTENTIONALLY.

Table of Contents

List of Illustrations

THIS PAGE IS BLANK INTENTIONALLY.

Acknowledgments

The authors would like to thank Ferne Friedman-Berg, Randy Sollenberger, and Carolina Zingale (ANG-E25) for the information that they provided on the ATWIT. The authors would also like to thank April Jackman (TASC, Inc.) and Eric Taylor (TG O'Brien, Inc.) for their support with editing.

THIS PAGE IS BLANK INTENTIONALLY.

Executive Summary

This paper reviews the Air Traffic Workload Input Technique (ATWIT). Stein (1985) developed the ATWIT as a 10-point scale to assess air traffic controller workload. Since the initial development, other researchers have implemented a 7-point version of this scale, but an empirical analysis to validate this approach is absent from the literature. The purpose of this paper is to provide a preliminary basis for recommending one version of the ATWIT over another.

A review of scale development provides a foundation for the analysis. Scales with greater reliability benefit the researcher by (a) increasing a measured effect size, (b) making it easier for researchers to conduct experiments, and (c) reducing costs associated with experimentation. The authors hypothesize that the scale size of the ATWIT may impact its reliability.

A preliminary literature review identified 15 studies for inclusion in a meta-analysis. The authors examined the impact of ATWIT scale size on correlation coefficients measuring the relationship between the aircraft count and ATWIT ratings. Results indicated that the average strength of the correlation was significantly greater for the 10-point version of the ATWIT compared to the 7-point version. This implies that the 10-point version of the scale may be more reliable and, therefore, superior to the 7-point version. Due to the preliminary nature of this analysis, the authors recommend additional research to examine and control for the effects of potential confounds.

THIS PAGE IS BLANK INTENTIONALLY.

1. INTRODUCTION

In the context of aviation and human factors research, Vidulich and Tsang (2012) defined *workload* as the mental resources expended to complete the demands of a task. The assessment of workload has been a vital component of research for the Federal Aviation Administration (FAA) for decades. Various studies have shown that controller workload is associated with inflight communication (Manning, Mills, Fox, Pfleiderer, & Mogilka, 2002; Stein, 1985) and number of aircraft to control (Ahlstrom & Friedman-Berg, 2006; Crutchfield & Rosenberg, 2007; Hah & Willems, 2008; Hah, Willems, & Phillips, 2006; Lee, 2005; Manning, Mills, Fox, Pfleiderer, & Mogilka, 2001a, 2001b; Manning et al., 2002; Rantanen, 2004; Sollenberger, La Due, Carver, & Heinze, 1997; Sollenberger & Stein, 1995; Stein, 1985; Willems, Allen, & Stein, 1999; Willems & Heiney, 2002; Yang, Rantanen, & Zhang, 2010). In addition, studies have shown that controller workload is associated with weather-related stressors (Ahlstrom & Friedman-Berg, 2006), situational awareness (Endsley & Rodgers, 1997), and poor runway approach (Stein, 1989). Workload is also an effective tool to support the development of systems and standards (Allendoerfer, Galushka, Mogford, 2000; McAnulty, Zingale, & Willems, 2005; Sollenberger & Hale, 2011). Therefore, it would benefit the FAA to standardize and optimize the research tools that it uses to assess workload.

A common tool for workload assessment used at the FAA William J. Hughes Technical Center (WJHTC) is the Air Traffic Workload Input Technique (ATWIT; Stein, 1985). The ATWIT scale provides a real-time assessment of workload in air traffic control (ATC) simulations (Ahlstrom & Friedman-Berg, 2006; Endsley, Mogford, Allendoerfer, Snyder, & Stein, 1997; Manning et al., 2002; McAnulty et al., 2005; Pfleiderer, 2005; Sollenberger & Hale, 2011; Sollenberger et al., 1997; Sollenberger & Stein, 1995; Willems et al., 1999; Willems & Heiney, 2002). However, some applications of the ATWIT have deviated from the original scale. Namely, some versions of the ATWIT contained 7 points and others contained 10 points.

Seemingly minor aspects of a scale may have a notable impact on its effectiveness (Nunnally & Bernstein, 1994). Because scales may be optimized with a variety of different scale sizes (Bendig, 1954a, 1954b; Benjamin, Tullis, & Lee, 2013; Cicchetti, Showalter, & Tyrer, 1985; Jenkins & Taber, 1977; Preston & Colman, 2000), it can be difficult to determine the ideal size of a scale. Chang (1994) argued that the ideal scale size will depend on the specific research context. To maximize the usefulness of the ATWIT, there is a need (a) to review studies that have used the ATWIT scale and (b) to identify how differences in the scale size may have impacted the ATWIT assessment of workload.

2. RELATED WORK

2.1 The Impact of Reliability on Effect Size

Scale reliability is the degree to which a measured score correlates with the true score that the scale intends to assess (Allen & Yen, 2002). Figure 1 illustrates this concept for two hypothetical scales. A reliable scale (in this case Scale A) has high degree of overlap with the true score, which increases the ability to accurately assess the construct of interest. An unreliable scale has much less overlap, meaning a high degree of error (in this case Scale B). Error leads to greater variability across measurements and increases the likelihood of reporting inaccurate relationships. The primary goal for developing an effective scale is to maximize reliability and decrease error.

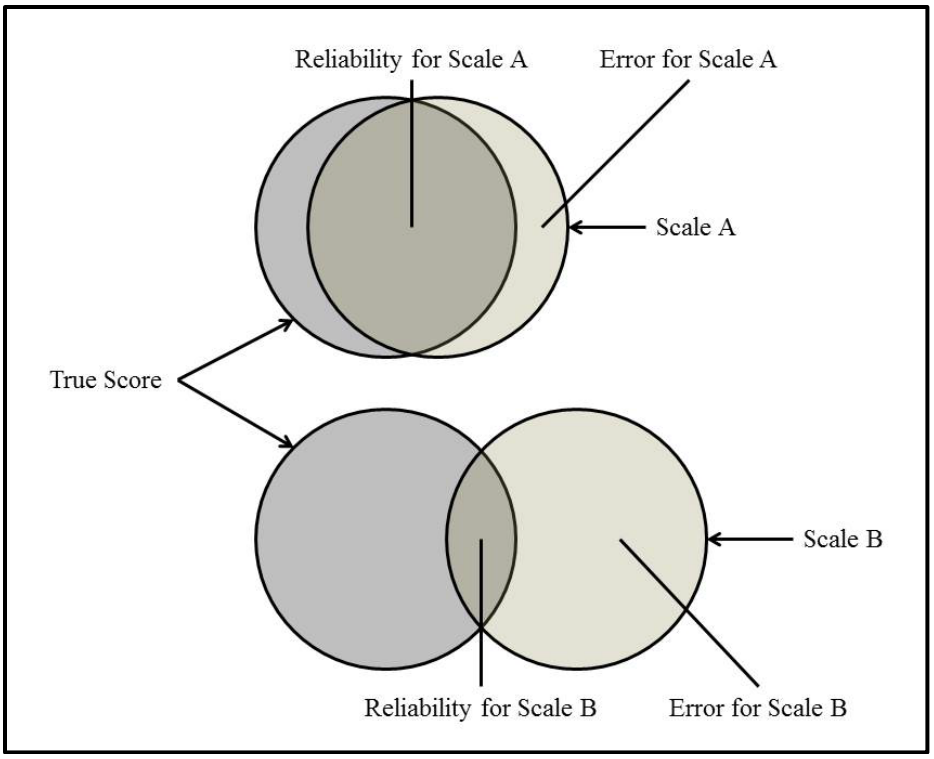

Figure 1. Venn diagram for illustrating reliability of assessment of a true score by two metrics.

One important aspect of reliability is its impact on effect size (Hunter & Schmidt, 2004; Nunnally & Bernstein, 1994; Shadish, Cook, & Campbell, 2002). Reliable scales generally yield higher and more accurate estimates of effect size than less reliable scales. Conversely, an unreliable scale produces more variation (e.g., measurement error), and this decreases effect size. Central to this review is the notion that an increase in scale reliability is associated with an increase in observed effect size.

Increasing the strength of an effect via improved reliability has practical benefits (see Figure 2). Shadish et al. (2002) noted that higher reliability increases statistical power. When using a less reliable scale, other steps must be taken to increase statistical power, such as increasing sample size. However, recruiting more participants for large-scale ATC studies is not always feasible due to budget constraints and the availability of active controllers. Low reliability also increases the variability of results across experiments (Borenstein, Hedges, Higgins, & Rothstein, 2009; Hunter & Schmidt, 2004). This variability may lead to scientific disagreements regarding the generalizability of findings, prompting additional experiments or meta-analyses to resolve these conflicts, which can also increase costs. For these reasons, maximizing reliability can save time, money, and effort.

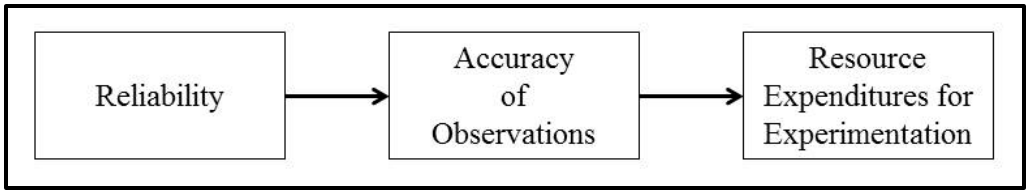

Figure 2. Practical impact of reliability on resource expenditures for experimentation.

2.2 Scaling Size

The size of a scale refers to the number of points that provide a measurement. Scale size is important, because it impacts reliability and effect sizes (Benjamin et al., 2013; Chang, 1994; Krosnick, Holbrook, & Visser, 2006; Preston & Colman, 2000). Different studies have supported scales of varying size, ranging from 2 points to 10 points (Bendig, 1954a, 1954b; Benjamin et al., 2013; Cicchetti et al., 1985; Jenkins & Taber, 1977; Preston & Colman, 2000). The preferred size for a given application depends on the context in which the scale is being used (Chang, 1994). For instance, Krosnick et al. (2006) argued that bipolar scales (e.g., ranging from agree to disagree with a neutral mid-point) were optimized with a 7-point scale, whereas unipolar scales (e.g., ranging from no usage to extreme usage) were optimized with a 5-point scale.

The literature presents no simple answer to what the ideal size of the ATWIT scale may be. Instead, we should consider the unique characteristics of the ATWIT and ask whether previous studies using the ATWIT favor one scale size over another.

2.3 The Air Traffic Workload Input Technique

Stein (1985) designed the ATWIT scale to collect real-time assessments of workload by requesting ratings from a participant at regular intervals during a task. Ratings with the ATWIT scale show a strong relationship with taskload factors, such as the number of aircraft (Ahlstrom & Friedman-Berg, 2006; Hah et al., 2006; Manning et al., 2002; Sollenberger et al., 1997; Sollenberger & Stein, 1995; Stein, 1985), supporting its use as a valid measure of workload. The original ATWIT was a 10-point scale that anchored across four self-perceived levels of error:

1. All tasks complete (i.e., points 1 and 2),
2. Little chance of error (i.e., points 3 to 5),
3. Some chance of error (i.e., points 6 to 8), and
4. Tasks are likely to be missed (i.e., points 9 and 10).

Later studies adopted a modified version of the ATWIT with only 7-point scales (Manning et al., 2002; Rantanen, 2004), which preserved the high vs. low distinction but removed the behavioral anchors. One argument for this involved standardizing the number scaling points across measures (Allendoerfer & Galushka, 1999), which was intended to minimize confusion and error by participants. Theoretically, there are reasons to favor both scale versions. Basic research favors the use of either 7-point scales (Cicchetti et al., 1985) or 5-point scales (Krosnick et al., 2006)—though differences in the research context may favor different scale sizes (Chang, 1994; Pasek & Krosnick, 2010), including larger scales of up to 10 points.

An important property of ATC research is that controllers tend to use a small portion of the workload scale—from low to medium levels of workload (McAnulty et al., 2005; Yang et al., 2010). Presumably, this is due to the use of active air traffic controllers as research participants who are highly trained experts with recent experience tolerating high traffic loads and stressful work environments. Participants in typical ATC studies use only a small portion of the scale—thus, researchers may benefit from more points at the lower end of the scale to ensure adequate resolution when measuring controller workload.

3. METHODOLOGY

In general, more reliable scales will produce stronger effect sizes. To examine whether some scales may be more reliable than others, the authors conducted a meta-analysis using studies with varying scale sizes. The meta-analysis focuses on the relationship between the number of aircraft in a controller's sector and workload ratings using the ATWIT. The analysis uses measures of this relationship to compare effects sizes with a 7-point ATWIT to those with a 10-point ATWIT.

3.1 Search & Inclusion Criteria

The authors conducted a literature search to find relevant studies. The databases for this search included the (a) FAA database, (b) Google Scholar, (c) HFES Proceedings, and (d) IEEE Xplore. Search terms included the (a) ATWIT and (b) Workload Assessment Keypad. Google scholar's "cited by" feature enabled a search for research that cited Stein's (1985) original paper. The search also looked for unpublished data. This literature search located 497 papers (K = 497).

Further review of papers focused on elements of inclusion. This began by identifying duplicate publications and non-empirical research. Of the studies that used the ATWIT for experimentation, this analysis only included studies that were designed to examine the effects of aircraft count. The process continued by selecting studies that used a 7-point scale or a 10-point scale to assess workload. This review identified a total of 15 studies (K = 15), which included 5 studies with a 7-point scale and 10 studies with a 10-point scale. Figure 3 shows a PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analysis) diagram to illustrate this process.

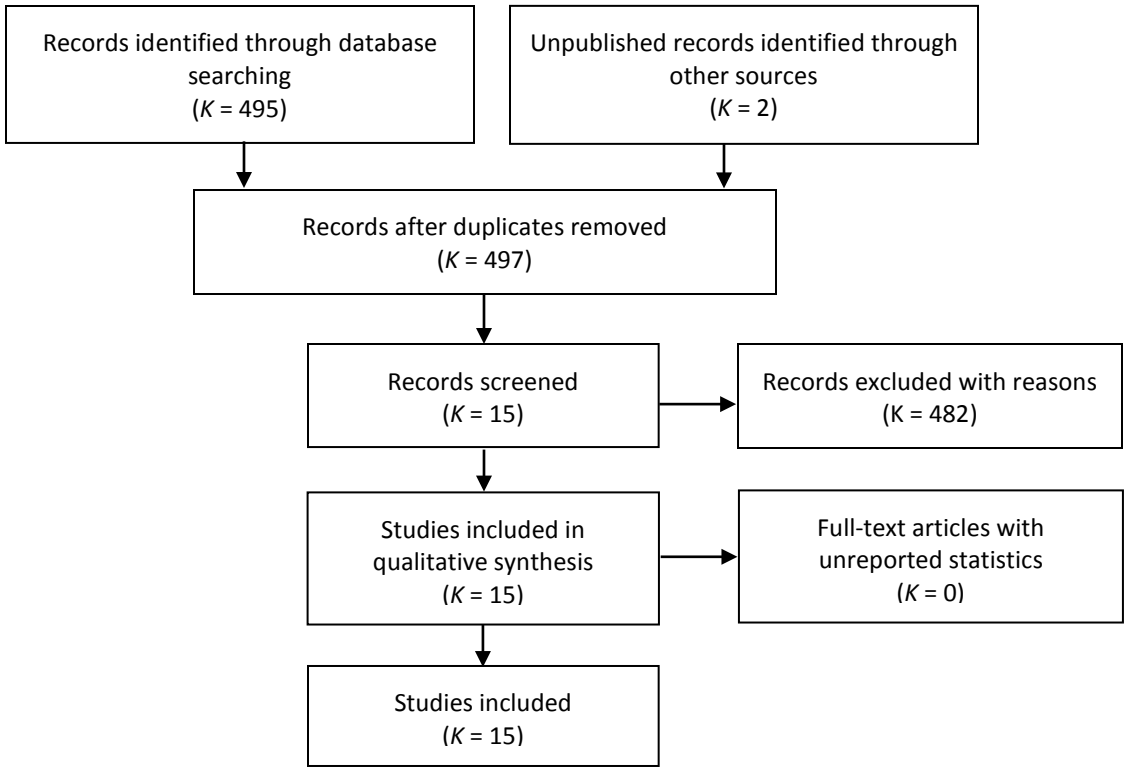

Figure 3. Preferred reporting items for systematic reviews and meta-analysis (PRISMA) flow diagram.

3.2 Data Extraction

The meta-analysis used correlation coefficients as the primary metric. When the study did not provide correlation coefficients, the authors used transformations of other statistics to obtain correlation coefficients.

3.3 Design & Analysis

The authors used the Hunter and Schmidt (2004) method of meta-analysis to aggregate data. This method provided uncorrected averages, standard deviations, and 95% confidence intervals for the 7-point scale and 10-point scale versions of the ATWIT. The Q statistics provided an assessment of moderation for this analysis. The Q statistic is a test of heterogeneity that measures the weighted sum of squares and uses a chi-squared distribution to determine significance.

4. RESULTS

A search for relevant literature identified 15 studies for this analysis, including 10 studies with a 10-point version of the ATWIT and five studies with a 7-point version of the ATWIT. Table 1 and Table 2 provide a summary of this data, which indicates that the average correlation for the 10-point version of the ATWIT ($r = .847$) was greater than the average correlation for the 7-point version of the ATWIT ($r = .733$). A Q-statistic provided an assessment of moderation and indicated that the difference between the 7-point and 10-point versions of the ATWIT was statistically significant ($Q = 5.41$, $p < 0.05$). The Q statistic also provided an assessment of unexpected confounds within the averages for the 7-point and 10-point versions of the ATWIT, but these were not statistically significant.

Table 1. Effect Size and Participant Number by Study

Study	Scale version	Correlation	Participant number
Lee (2005)	7-point	.711	22
Manning et al. (2001a)	7-point	.800	16
Manning et al. (2001b)	7-point	.699	16
Manning et al. (2002)	7-point	.752	16
Rantanen (2004)	7-point	.709	16
Ahlstrom & Friedman-Berg (2006)	10-point	.866	6
Crutchfield & Rosenberg (2007)	10-point	.828	2
Hah et al. (2006)	10-point	.721	16
Hah & Willems (2008)	10-point	.811	12
Sollenberger & Stein (1995)	10-point	.953	16
Sollenberger et al. (1997)	10-point	.949	16
Stein (1985)	10-point	.899	12
Willems et al. (1999)	10-point	.790	10
Willems & Heiney (2002)	10-point	.938	16
Yang et al. (2010)	10-point	.767	31

Table 2. Summary Statistics of the Association Between the ATWIT and the Number of Aircraft During an ATC Task for 7-Point and 10-Point Versions of the ATWIT

ATWIT Scaling	*N*	*K*	*r*	SD_r	$L\text{-}CI_r$ 95%	$U\text{-}CI_r$ 95%	*Q*
10-point scale	137	10	.847	.080	.798	.896	11.94
7-point scale	86	5	.733	.024	.712	.753	0.51

Note. N = total number of participants across studies; K = number of studies; r = mean observed correlation; SD_r = standard deviation for r; $L\text{-}CI_r$ 95% = lower limit of the 95% confidence interval for r; $U\text{-}CI_r$ 95% = upper limit of the 95% confidence interval for r; Q = moderator statistic, where $df = K - 1$ on a chi-squared distribution.

5. DISCUSSION

5.1 Recommendations

A meta-analysis of 15 previous ATC studies conducted at WJHTC indicated that the 10-point version of the ATWIT tended to yield stronger effect sizes. Based on this finding, the 10-point version of the ATWIT appears to provide a superior assessment of ATC workload. This is likely attributed to the fact that participants in ATC studies are expert air traffic controllers who require more points to differentiate between lower levels of workload.

5.2 Limitations

The above meta-analysis is a preliminary assessment of the relationship between the ATWIT scale size and scale reliability. The meta-analysis did not control for methodological differences in the 7-point and 10-point studies, including other taskload manipulations (weather, extreme events, etc.) or administration of the ATWIT (e.g., workload assessment keypad vs. touch screen). Future analyses of the ATWIT should investigate the degree to which these factors alter this analysis.

5.3 Concluding Remarks

This paper provides a preliminary assessment of the ATWIT scale size, and the results support the use of the 10-point version of the ATWIT. These results were attributed to the finding that expert air traffic controllers, who are used in ATC research studies, tend to report lower levels of workload. The 10-point version of the ATWIT provides more points at the lower end of the ATWIT; therefore, it may provide a more reliable assessment of controller workload. Future efforts should evaluate ATWIT scale size while controlling for potential confounds, such as taskload factors.

References

References marked with an asterisk indicate studies included in the meta-analysis.

*Ahlstrom, U., & Friedman-Berg, F. (2006). Using eye movement activity as a correlate of cognitive workload. *International Journal of Industrial Ergonomics, 36*, 623–636. doi:10.1016/j.ergon.2006.04.002

Allen, M., & Yen, W. (2002). *Introduction to measurement theory*. Long Grove, IL: Waveland Press, Inc.

Allendoerfer, K., & Galushka, J. (1999). *Air traffic control system baseline methodology guide* (DOT/FAA/CT-TN99/15). Atlantic City International Airport, NJ: William J. Hughes Technical Center.

Allendoerfer, K., Galushka, J., & Mogford, R. (2000). *Display system replacement baseline research report* (DOT/FAA/CT-TN00/31). Atlantic City International Airport, NJ: William J. Hughes Technical Center.

Bendig, A. (1954a). Reliability and the number of rating scale categories. *Journal of Applied Psychology, 38*, 38–40. doi:10.1037/h0055647

Bendig, A. (1954b). Reliability of short rating scales and the heterogeneity of rated stimuli. *Journal of Applied Psychology, 38*, 167–170. doi:10.1037/h0059072

Benjamin, A., Tullis, J., & Lee, J. (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*(5), 1601–1608. doi:10.1037/a0031849

Borenstein, M., Hedges, L., Higgins, J., & Rothstein, H. (2009). *Introduction to meta-analysis*. Hoboken, NJ: John Wiley & Sons.

Chang, L. (1994). A psychometric evaluation of 4-point and 6-point likert-type scales in relation to reliability and validity. *Applied Psychological Measurement, 18*(3), 205–215. doi:10.1177/014662169401800302

Cicchetti, D., Showalter, D., & Tyrer, P. (1985). The effect of number of rating scale categories on levels of interrater reliability: A Monte Carlo investigation. *Applied Psychological Measurement, 9*, 31–36. doi:10.1177/014662168500900103

*Crutchfield, J., & Rosenberg, C. (2007). *Predicting subjective workload ratings: A comparison and synthesis of operational and theoretical models* (DOT/FAA/AM-07/6). Washington, DC: Office of Aerospace Medicine.

Endsley, M., Mogford, R., Allendoerfer, K., Snyder, M., & Stein, E. (1997). *Effect of free flight conditions on controller performance, workload, and situation awareness* (DOT/FAA/CT-TN97/12). Atlantic City International Airport, NJ: William J. Hughes Technical Center.

Endsley, M., & Rodgers, M. (1997). *Distribution of attention, situation awareness, and workload in a passive air traffic control task: Implications for operational errors and automation* (DOT/FAA/AM-97/13). Oklahoma City, OK: FAA Civil Aeromedical Institute.

*Hah, S., & Willems, B. (2008). The relationship between aircraft count and controller workload in different en route workstation systems. In *Proceedings of the 52nd Annual Meeting of the Human Factors and Ergonomics Society, 52*, 44–48.

*Hah, S., Willems, B., & Phillips, R. (2006). The effect of air traffic increase on controller workload. In *Proceedings of the 50th Annual Meeting of the Human Factors and Ergonomics Society*, 50–54.

Hunter, J., & Schmidt, F. (2004). *Methods of meta-analysis: Correcting error and bias in research findings.* Thousand Oaks, CA: Sage Publications, Inc.

Jenkins, G., & Taber, T. (1977). A monte carlo study of factors affecting three indices of composite scale reliability. *Journal of Applied Psychology, 29*, 66–68. doi:10.1037/0021-9010.62.4.392

Krosnick, J., Holbrook, A., & Visser, P. (2006). Optimizing brief assessments in research on the psychology of aging: A pragmatic approach to survey and self-report measurement. In L. Carstensen and C. Hartel (Eds.), *National research council (us) committee on aging frontiers in social psychology, personality, and adult development psychology* (pp. 231–239). Washington, DC: National Academies Press.

*Lee, P. (2005). A non-linear relationship between controller workload and traffic count. In *Proceedings of the 49th Annual Meeting of the Human Factors and Ergonomics Society, 49*, 1129–1133.

*Manning, C., Mills, S., Fox, C., Pfleiderer, E., & Mogilka, H. (2001a). *Investigating the validity of performance and objective workload evaluation research* (POWER; DOT/FAA/AM-01/10). Washington, DC: Office of Aerospace Medicine.

*Manning, C., Mills, S., Fox, C., Pfleiderer, E., & Mogilka, H. (2001b). The relationship between air traffic control communication events and measure of controller taskload and workload. In the *4th USA/Europe Air Traffic Management R&D Seminar*. Santa Fe, NM.

*Manning, C., Mills, S., Fox, C., Pfleiderer, E., & Mogilka, H. (2002). *Using air traffic control taskload measures and communication events to predict subjective workload* (DOT/FAA/AM-02/4).Washington, DC: Office of Aerospace Medicine.

McAnulty, D., Zingale, C., & Willems, B. (2005). Controller-in-the-loop evaluation of traffic management advisor (TMA) metering data format and location. In *Proceedings of the Mini-Conference of Human Factors in Complex Sociotechnical Systems, 5,* 1–5.

Nunnally, J., & Bernstein, I. (1994) *Psychometric theory* (3rd ed.). New York: McGraw-Hill, Inc.

Pasek, J., & Krosnick, J. (2010). Optimizing survey questionnaire design in political science: Insights from psychology. In J. Leighley (Eds.), *The Oxford handbook of American elections and political behavior* (pp. 27–50). Oxford, UK: Oxford University Press.

Pfleiderer, E. (2005). The good, the not-so-bad, and the ugly: Computer-detected altitude, heading, and speed changes in en route air traffic control. In *Proceedings of the Mini-Conference on Human Factors in Complex Sociotechnical Systems, 4,* 1–5.

Preston, C., & Colman, A. (2000). Optimal number of response categories in rating scales: Reliability, validity, discriminating power, and respondent preferences. *Acta Psychologica, 104*, 1–15. doi:10.1016/S0001-6918(99)00050-5

*Rantanen, E. (2004). *Development and validation of objective performance and workload measures in air traffic control* (AHFE-04-19/FAA-04-07). Savoy, IL: Aviation Human Factors Division Institute of Aviation.

Shadish, W. R., Cook, T.D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* New York: Houghton Mifflin Company.

Sollenberger, R., & Hale, M. (2011). Human-in-the-loop investigation of variable separation standards in the en route air traffic control environment. In *Proceedings of the 55th Annual Meeting of the Human Factors and Ergonomics Society*, 66–70.

*Sollenberger, R., La Due, J., Carver, B., & Heinze, A. (1997). *Human factors evaluation of vocoders for air traffic control environments phase II: ATC simulation* (DOT/FAA/CT-TN97/25). Atlantic City International Airport, NJ: Federal Aviation Administration Technical Center.

*Sollenberger, R., & Stein, E. (1995). *The effects of structured arrival and departure procedures on TRACON air traffic controller memory and situation awareness* (DOT/FAA/CT-TN95/27). Atlantic City International Airport, NJ: Federal Aviation Administration Technical Center.

*Stein, E. (1985). *Controller workload: An examination of workload probe* (DOT/FAA/CT-TN84/24). Atlantic City International Airport, NJ: Federal Aviation Administration Technical Center.

Stein, E. (1989). *Parallel approach separation and controller performance* (DOT/FAA/CT-TN89/50). Atlantic City International Airport, NJ: Federal Aviation Administration Technical Center.

Vidulich, M., & Tsang, P. (2012). Mental workload and situation awareness. In G. Salvendy (3rd eds.), *Handbook of human factors and ergonomics* (3rd ed., pp. 243–273). Hoboken, NJ: John Wiley & Sons, Inc.

*Willems, B., Allen, R., & Stein, E. (1999). *Air Traffic Control Specialist visual scanning II: Task load, visual noise, and intrusions into controlled airspace* (DOT/FAA/CT-TN99/23). Atlantic City International Airport, NJ: Federal Aviation Administration William J. Hughes Technical Center.

*Willems, B., & Heiney, M. (2002). *Decision support automation research in the en route air traffic control environment* (DOT/FAA/CT-TN02/10). Atlantic City International Airport, NJ: Federal Aviation Administration William J. Hughes Technical Center.

*Yang, J., Rantanen, E., & Zhang, K. (2010). The impact of time efficacy on air traffic controller situation awareness and mental workload. *The International Journal of Aviation Psychology, 20*(1), 74–91. doi:10.1080/10508410903416037