HOW BUSY IS TOO BUSY? VALIDATION OF THE DUTCH INTEGRATED WORKLOAD SCALE (IWS)

Robin Kramer BSc[1], Prof. Dr. Addie Johnson[2], Melcher Zeilstra MSc EurErg[3]

*[1]University of Groningen, Faculty of Mathematics and Natural Sciences,
9747 AG, Groningen, The Netherlands, E-mail: r.kramer.2@student.rug.nl
[2]University of Groningen, Faculty of Behavioural and Social Sciences,
9712 TS, Groningen, The Netherlands, E-mail: a.johnson@rug.nl
[3]Intergo, consultancy in human factors and ergonomics, P.O. Box 19218,
3501 DE, Utrecht, The Netherlands, E-mail: zeilstra@intergo.nl*

The Integrated Workload Scale (IWS) is a subjective scale for real-time workload assessment. The nine anchor points of the IWS—ranging from "*no workload*" to "*work too demanding*"—capture the multidimensionality of the concept of workload by incorporating items reflecting time and effort in addition to workload and demand. Although originally designed for research with train dispatchers in the United Kingdom, the IWS has since been used to measure workload in other countries. In one of these countries, the Netherlands, the IWS has been translated into Dutch for use with train dispatchers. The aim of the present study was to validate the Dutch translation of the IWS. Dutch students and train dispatchers and English-language students were asked to rate the individual items of the IWS, in Dutch and English, respectively, on a scale from 0 ("no workload at all") to 150 ("complete overload"). The mean ratings of items did not differ significantly between the groups, suggesting that the two versions of the IWS were interpreted similarly. Regression analyses showed that the scales were perceived as linear, with equidistant items. Additional, alternate Dutch items were also rated as possible substitutions for some of the original items but were not found to significantly improve the linearity of the scale. The strong similarities of the Dutch IWS to the original IWS – including its multidimensional nature and the equidistant items – as well as the fact that train dispatchers and students gave similar ratings on the Dutch IWS suggest that it can reliably be used to assess subjective workload.

Introduction

The work of a train dispatcher varies across the course of a day. Much of the time, the task is to monitor the status of the rail network, and, if necessary, resolve conflicts in train traffic. When everything is going as planned and no intervention is necessary, task demands are low. However, when many conflicts are encountered task demands increase, and may become so high that the dispatcher has difficulty keeping up with the task.

Extreme levels of workload, whether high or low, are associated with relatively poor performance (Brookhuis and De Waard, 2002). When workload is too low, people will have difficulty staying involved in the task. This can seriously impact situation awareness, that is, the operator's perception, interpretation and anticipation of the current situation (Endsley, 2013), and, in turn, affect performance (Onnasch, Wickens, Li and Manzey, 2014). Workload that is too high also affects performance, and can even affect the health of the operator if stress is prolonged (Spieker et al., 2002). Knowing how demanding a job is for a given operator is therefore of vital importance. Identifying when peaks and lows occur, and how operators perceive these moments, allows one to make changes that could moderate the level of workload. What is needed is a method tailored to the task of interest to accurately measure workload in the field as the load is experienced.

Subjective Assessment

A common method of workload measurement is subjective assessment, in which operators are asked to estimate their experienced level of workload. The NASA Task Load Index (NASA-TLX; Hart and Staveland, 1988) is a widely used scale of this type. This scale makes a distinction between six factors that all contribute to the perceived workload of an operator. These factors are (1) temporal demand, (2) mental demand, (3) physical demand, (4) performance, (5) effort and (6) frustration. Because using the NASA-TLX requires responding on six separate scales (as well making comparisons between the scales), it may divert attention from the primary task in much the same way as does a secondary task, which may, in turn, increase the workload that is being measured (Katidioti, Borst and Taatgen, 2014).

In an effort to simplify the measurement of mental workload, Zijlstra (1993) developed the Rating Scale Mental Effort (RSME), a one-dimensional scale that has been used in transportation environments. The RSME, which has been translated into several languages, including Indonesian (Widyanti, Johnson and De Waard, 2013), is a scale ranging from 0 to 150, on which nine anchor points are marked. The anchor points, which are marked at uneven intervals are descriptions of levels of perceived mental effort, such as *some effort* or *considerable effort*. All of the anchor points of the RSME describe a degree of effort; as such the RSME does not incorporate other dimensions of workload.

The Integrated Workload Scale (IWS; Pickup, Wilson, Norris, Mitchell and Morrisroe, 2005) is a relatively new scale that capitalizes on the simplicity of a one-dimensional scale without sacrificing the advantages of a multi-dimensional definition of workload. It incorporates several dimensions from the NASA-TLX, including temporal demand and effort, into a single scale. Instead of reporting workload by marking a line as in the RSME, workload is reported by indicating one of nine anchor points, which are colour coded and accompanied short descriptive texts (see Figure 1).

The nine anchor points of the IWS were selected, based on a rating procedure, from a pool of descriptors derived from descriptions of workload collected during interviews with and observation of train dispatchers (Pickup et al., 2005). The descriptors were based on dispatcher reports to ensure that the anchor points would contain terminology familiar to dispatchers. It was noted that dispatchers tended to refer to multiple dimensions of workload (e.g. time available, pressure and frustration; see also Ames and George (1993); this multidimensionality was preserved in the final scale.

The IWS is attractive because of its relative simplicity of use and the incorporation of a multi-dimensional definition of workload. Moreover, the IWS has been successfully applied in the past to measure workload in dispatchers working under different levels of automation (Balfe, Wilson, Sharples and Clarke, 2012). Balfe et al. showed, as expected, that during manual control workload is highest compared to two different automation systems. More interestingly, the IWS revealed a clear distinction between the two automation systems, and between moments of low and high workload. These factors led to the choice to adopt the IWS in the Netherlands for on-the-job investigation of workload in train dispatchers.

The Dutch version of the IWS used in previous research (Zeilstra, 2007) was the starting point for the current project. In a study by Wilms and Zeilstra (2013), the Dutch IWS was compared to TaskWeighing™, a tool that is used to estimate the amount of workload imposed by a task (Zeilstra, De Bruin, Van Der Weide; 2009). Promising, positive correlations were found between the results obtained using TaskWeighing™ and the Dutch IWS, but no direct comparison between the original and the Dutch IWS has yet been made. For our study, the Dutch version of the IWS used in previous research, which had been created by a team of Dutch human factors professionals was first back-translated by a native English-speaking human factors professional (the second author). Items for which the translation was not perfect were discussed with the original translators and amended. The English IWS items and their Dutch translations are shown in Table 1. A number of alternate items for which multiple translations were possible were retained to allow for possible substitutions after later testing (see Table 2).

Table 1: English and Dutch items of the IWS

Colour		Item	Description
Blue	English	Not demanding	Work is not demanding at all
	Dutch	Niet belastend	Het werk is helemaal niet belastend
Lilac	English	Minimal effort	Minimal effort required to keep on top of situation
	Dutch	Minimale inspanning	Er is minimale inspanning nodig om de situatie onder controle te houden
Light Blue	English	Some spare time	Active with plenty of time available to complete less essential jobs
	Dutch	Enige tijd over	Ik ben actief maar heb enige tijd over om minder belangrijke taken te doen
Blue-Grey	English	Moderate Effort	Work demanding but manageable with moderate effort
	Dutch	Matige inspanning	Het werk is belastend maar kan worden gedaan met matige inspanning
Azure	English	Moderate Pressure	Moderate pressure, work is manageable
	Dutch	Gemiddeld druk	Gemiddeld druk, het werk kan worden gedaan
Green	English	Very busy	Very busy but still able to do job
	Dutch	Erg druk	Erg druk maar nog wel in staat de taak uit te voeren
Yellow	English	Extreme effort	Extreme effort and concentration necessary to ensure everything gets done
	Dutch	Zeer veel inspanning	Zeer veel inspanning en concentratie nodig om zeker te zijn dat alles gebeurt
Orange	English	Struggling to keep up	Very high level of effort and demand, struggling to keep up with everything
	Dutch	Moeite om het werk bij te houden	Zeer veel inzet en inspanning nodig, moeite om het werk bij te houden
Red	English	Work too Demanding	Work too demanding – complex or multiple problems to deal with and even with very high levels of effort it is unmanageable.
	Dutch	Te belastend	Het werk is te belastend en complex, of te veel problemen. Zelfs met zeer veel inzet is het werk niet te doen

Table 2: Alternate Dutch translations

Original Item	Dutch Translation	Alternatives
Some spare time	Enige tijd over	Lichte inspanning
Moderate Effort	Matige inspanning	Redelijke inspanning
Very busy	Erg druk	Behoorlijk druk Behoorlijk inspannend Enorm druk Enorm inspannend
Struggling to keep up	Moeite om het werk bij te houden	Extreme inspanning
Work too Demanding	Te belastend	Overbelastend

In order to further verify that the English and Dutch versions of the IWS were comparable, Dutch and English-language students and train dispatchers were asked to rate the items of the IWS according to the workload (*werkbelasting*) conveyed by the items. Students were compared because of the availability of separate Dutch and English-language student populations. Train dispatchers were included to ensure that the students' understanding of the items was similar to that of the main population of interest.

Method

A total of 125 subjects participated in the study. Of the University of Groningen student participants, 58 were students following the English-language psychology program (30 female; age = 21.4 ± 2.2 years) and 48 were Dutch students from different programs (28 female; age = 21.8 ± 1.9 years). Although the language of instruction is English for the English-language group, English is not the first language for 87.9% of the students. The 19 Dutch train dispatchers (1 female; age = 43.2 ± 11.4 years) all worked at the Groningen dispatch centre. Of the train dispatchers, the level of education was university (26.3%), mid-vocational (42.1%) and secondary school (31.6%) as opposed to all university in the student groups. All participants gave informed consent; 82 students received partial course credit for participation. Dutch students and train dispatchers rated the Dutch translation of the IWS items and alternate items (see Table 2); English-speaking students rated the original English items only.

Data from 15 participants were excluded due to incompleteness of the data (i.e. failure to rate four or more of the IWS items). Ratings of one participant who failed to rate four alternate items and of three participants who failed to rate one item were replaced with the mean rating of that item. A further 12 participants were excluded due to suspected noncompliance with task instructions as indicated by giving consistently low (<70) ratings to high workload items or consistently high (>80) ratings to low workload items. In one case a rating considered to be a typing error was corrected. A value of 7 was reported where

“70” would have been expected and consistent with all other responses of the subject. This was done by replacing it with the mean rating of that item.

A magnitude estimation (ME; see Meek, Sennott-Miller, and Ferketich, 1992) procedure was used in which participants were to assign a value ranging from 0 (“no workload at all”) to 150 (“complete overload”) to each. For the English version of the task, the nine original items were presented in random order. For the Dutch version of the task, a block with the nine items corresponding to the IWS, presented in random order, was followed by a block containing the alternate translations, also in random order. The alternate items were presented after the ISW items so as not to affect estimates of the original Dutch items.

Results

We first compared the Dutch and English student ratings of the IWS items (see Figure 1) in order to determine if there were significant differences in how the two groups interpreted the items. Train dispatchers were not included in this analysis because of large differences in group size. A 2 (group) x 9 (item) repeated measures ANOVA was performed on the nine items of the IWS rated by the English-speaking and Dutch students. As expected, a main effect of item was found ($F(8, 672) = 517.35$, $p < 0.001$), such that ratings increased with increasing load. However, there was no significant difference between groups ($F(1,84) = 0.051$, $p = 0.824$) and no Item x Group interaction ($F(8, 672) = 0.97$, $p = 0.460$).

Because the IWS was designed to have equidistant items (Pickup et al., 2005), we next tested whether ratings were linear with respect to the item numbering. Three simple linear regression lines were calculated for the three different groups (train dispatchers, and English-speaking and Dutch students). The regression line for the English-speaking students had a slope, $\beta = 15.63$ ($t(438) = 36.8$, $p < 0.001$) and $R^2 = 0.778$ ($F(1,439) = 1536.45$, $p < 0.001$). Similar results were found for the Dutch students ($\beta = 15.88$, $t(330) = 36.82$, $p < 0.001$; $R^2 = 0.804$, $F(1,331) = 1355.75$, $p < 0.001$) and train dispatchers ($\beta = 15.18$, $t(105) = 22.17$, $p < 0.001$; $R^2 = 0.823$, $F(1,106) = 491.70$, $p < 0.001$). The high R^2 values indicate a high degree of linearity as a function of item number, or, in other words, that the items are rated as being approximately equidistant by each of the groups. The means for each item and regression lines as a function of group are shown in Figure 1.

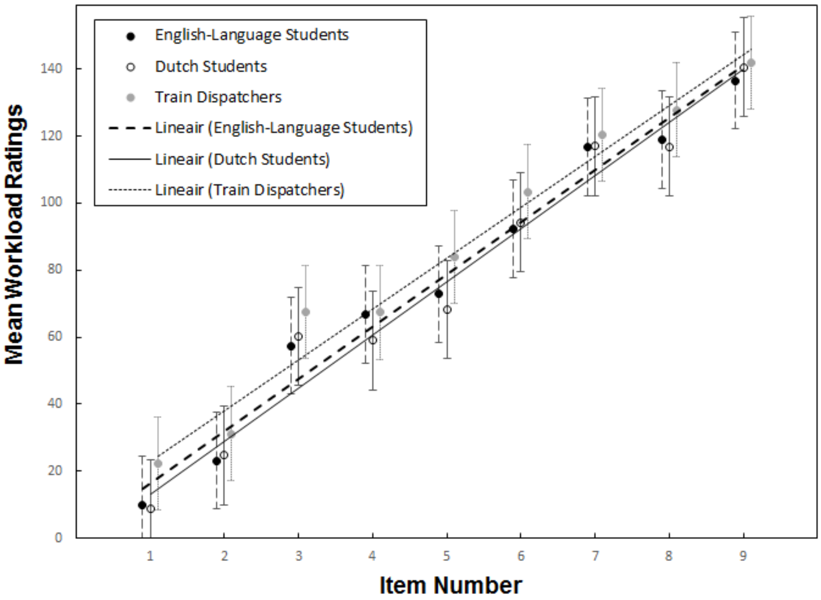

Figure 1: Mean workload ratings and regression lines for Dutch (Dutch students and Train Dispatchers) and English (English-language students) IWS items. The items are ordered from *No workload* (1) to *Work too demanding* (9). The error bars depict 95% confidence intervals.

Evaluating the alternate items

Although the Dutch IWS items chosen for testing fared well both in comparison to the English IWS and in terms of variance accounted for in a regression model, we still examined our alternate items to demine whether it was possible to create an even better instrument. The ratings given to the alternate items are in Table 3. Items that significantly differed from the original and were closer to the regression line were selected for further analysis. For the train dispatchers, none of the alternate items satisfied these conditions. For the students *"lichte inspanning"* ($m_{difference}$ = -9.9, t(36) = -2.419, p < .05) and *"extreme inspanning"* ($m_{difference}$ = 8.1, t(36) = 2.662, p < .05) were significantly closer to the regression line than the original Dutch items. These items were substituted for the original items and the resulting alternate Dutch IWS was tested using simple linear regression with both the Dutch student and train dispatcher data. For the students, β = 16.65 (t(331) = 40.42, p < 0.001) and R^2 = 0.831 (F(1,331) = 1633.39, p < 0.001). For the train dispatchers, β = 15.61 (t(106) = 23.85, p < 0.001) and R^2 = 0.843 (F(1,106) = 568.74, p < 0.001).

To test whether this alternate scale could be considered an improvement over the original Dutch IWS we compared the original and alternate Dutch versions of the IWS by computing the fit of the ratings to the regression line of the original English IWS, for each scale separately. The fit for the original Dutch IWS gave an R^2 of 0.813 (F(1,331) = 1442.96, p < 0.001) for Dutch students and 0.771

(F(1,106) = 359.64, $p < 0.001$) for the train dispatchers. For the alternate scale, the $R^2 = 0.826$ (F(1,331) = 1579.89, $p < 0.001$) for the Dutch students and $R^2 = 0.800$ (F(1,106) = 427.67, $p < 0.001$) for the train dispatchers. The R^2 values of the alternate scale were numerically higher. To see if these differences were significant, we converted the R^2 values to correlations. Z-tests performed on the correlation coefficients showed that the improvement across the versions was not significant for Dutch students ($r_1 = 0.902$, $r_2 = 0.909$, $p = 0.609$), nor was it for train dispatchers ($r_1 = 0.878$, $r_2 = 0.894$, $p = 0.581$).

Table 3: Mean estimates (and SD) of the alternate Dutch translations

Alternate item	Original item	Dutch students	Train dispatcher
Lichte inspanning	Enige tijd over	50.3 (20.1)	53.8 (20.3)
Redelijke inspanning	Matige inspanning	70.6 (21.4)	88.4 (19.1)
Behoorlijk druk	Erg druk	92.4 (20.0)	96.2 (22.6)
Enorm druk	Erg druk	108.7 (16.8)	112.6 (14.9)
Behoorlijk inspannend	Erg druk	93.7 (22.8)	102.5 (21.2)
Enorm inspannend	Erg druk	105.4 (21.1)	113.4 (19.9)
Extreme inspanning	Moeite om het werk bij te houden	125.0 (18.4)	127.3 (13.9)
Overbelastend	Te belastend	141.9 (13.2)	144.8 (7.9)

Discussion

We conducted our study with two versions of the same instrument: an English and a Dutch version of the IWS. The Dutch IWS was the result of translating the English IWS, but this translation had yet to be validated. Therefore, we had students and Dutch train dispatchers perform a ME task on the items of the scale. The results of the ME task showed a considerable consistency in how the two items from the scale were rated, indicating that the Dutch translation of the IWS is very similar to its English counterpart. Moreover, both versions of the instrument showed evidence of roughly equal spacing of the anchor points of the workload scale.

Because two of the anchor points of the translated, Dutch scale showed some deviation from linearity (i.e. lack of equal spacing on the scale), an attempt was made to improve the scale by replacing these items with alternate items. The alternate items did marginally increase the variance accounted for by a linear regression, but the improvement to the scale was not significant. Another factor that should be considered in making any changes to the IWS would be the multidimensionality of the scale and the desirability of preserving that. In our case, the alternate translations that gave a numerically better fit than the translations originally chosen as being the closest to the original IWS were unidimensional in that they both referred to "effort" rather than other dimensions of load, as did the terms they would have replaced. This could be an important concern for translators of such instruments into other languages: It is important to balance the conceptual nature of the instrument with the psychometric properties of the scale.

The deviations from linearity found in the Dutch IWS were also found in the English IWS. The finding that the items of the English IWS are also not evenly spaced could, in part, be explained by the fact that English is not the first language of most of the English-language students who participated in our study. Subtle differences between words such as "*pressure*" and "*busy*", might not have been understood by some participants. During testing whether the alternate items improved the Dutch IWS. However, the level of proficiency required for entrance into the English-language program makes it unlikely that this would have had a large impact.

The fact that no significant differences between the Dutch students' and the train dispatchers' responses to the Dutch IWS were found, serves as an indication that they both have a similar understanding of the scale. Together with the fact that the train dispatchers were significantly older than the students and that many of the dispatchers had a lower level of education, this suggests that the IWS can reliably be used among different populations and, by extension, in fields other than that of train dispatching.

Conclusion

Comparison of the Dutch IWS and the original, English version of the IWS by means of magnitude estimates, revealed great overlap in how the two scales are viewed. The Dutch translation is now shown to be a valid measure for the subjective assessment of workload. The items are recognizable descriptions of workload for both students and train dispatchers—two groups that have very different backgrounds. The robustness of the IWS across user populations suggests a broad field of application.

Acknowledgements

Special thanks go out to ProRail train dispatchers of the Groningen Train Dispatching Centre for their cooperation in this research.

References

Ames, L. L., and George, E. J. (1993). *Revision and verification of a seven-point workload estimate scale* (No. AFFTC-TIM-93-01). AIR FORCE TEST CENTER EDWARDS AFB CA.

Balfe, N., Wilson, J. R., Sharples, S., and Clarke, T. (2012). Effects of level of signalling automation on workload and performance. *Wilson, JR, Mills, A., Clarke, T., Rajan, J., Dadashi, N.(eds.)*, 404-411.

Brookhuis, K. A., and De Waard, D. (2002). On the assessment of (mental) workload and other subjective qualifications. *Ergonomics*, *45*(14), 1026-1030.

Endsley, M. R., and Kiris, E. O. (1995). The out-of-the-loop performance problemand levelofcontrol in automation. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *37*(2), 381-394.

Endsley, M.R. (2013). Situation awareness. In J.D. Lee and A. Kirlik (eds.), *Oxford handbook of cognitive engineering*, Oxford University Press, Oxford, UK, 88-108

Hart, S. G., and Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. *Advances in psychology*, *52*, 139-183.

Katidioti, I., Borst, J. P., and Taatgen, N. A. (2014). What happens when we switch tasks: Pupil dilation in multitasking. *Journal of experimental psychology: applied*, 20(4), 380.

Meek, P. M., Sennott-Miller, L., and Ferketich, S. L. (1992). Focus on psychometrics scaling stimuli with magnitude estimation. *Research in nursing & health*, *15*(1), 77-81.

Onnasch, L., Wickens, C. D., Li, H., and Manzey, D. (2013). Human Performance Consequences of Stages and Levels of Automation An Integrated Meta-Analysis. *Human Factors: The Journal of the Human Factors and Ergonomics Society*.

Pickup, L., Wilson, J. R., Norris, B. J., Mitchell, L., and Morrisroe, G. (2005). The Integrated Workload Scale (IWS): a new self-report tool to assess railway signaller workload. *Applied Ergonomics*, *36*(6), 681-693.

Reason, J. (1990). Human error. Cambridge university press.

Spieker, L. E., Hürlimann, D., Ruschitzka, F., Corti, R., Enseleit, F., Shaw, S., ... and Noll, G. (2002). Mental stress induces prolonged endothelial dysfunction via endothelin-A receptors. Circulation, 105(24), 2817-2820.

Widyanti, A., Johnson, A., and De Waard, D. (2013). Adaptation of the rating scale mental effort (RSME) for use in Indonesia. *International Journal of Industrial Ergonomics*, *43*(1), 70-76.

Wilms, M. S., and Zeilstra, M. P. (2013). Subjective mental workload of Dutch train dispatchers: Validation of IWS in a practical setting. In *4th International Conference on Rail Human Factor*, 641-650.

Zeilstra, M. P. (2007). *Normering IWS*. Intergo, Utrecht, confidential.

Zeilstra, M. P., Bruijn, D., and Van der Weide, R. 2009, *Development and implementation of a predictive tool for optimizing workload of train dispatchers*, Intergo, Utrecht, confidential.

Zijlstra, F. R. H. (1993). *Efficiency in work behaviour: A design approach for modern tools*. TU Delft, Delft University of Technology.