Development and Validation of a Pictorial Scale for Tense Activation in Experience Sampling Methodology

Tim Odolphij

Studentnumber: s3680266

Department of Psychology

Rijksuniversiteit Groningen

Master Thesis (24/25)

Instructor/supervisor: Dr. Oliver Weigelt

Second assessor: Dr. Eric Rietzschel

6 January 2025

Abstract

The aim of this study is to develop and validate a pictorial scale for measuring tense activation in Experience Sampling Methodology (ESM). Tense activation, characterized by heightened feelings of anxiety, nervousness, and being on edge, along with possible physiological stress responses, currently lacks measurement instruments. The proposed scale uses a rubber-band metaphor that visually represents varying levels of tense activation, aiming to enhance participant engagement and minimize response fatigue. Two studies were conducted: the first assessed content validity through expert evaluations, revealing mixed results, particularly in terms of comprehensiveness. The second study examined the scale's convergent, discriminant, and criterion-related validity in an ESM context. Evidence was found for convergent validity, discriminant validity, and criterion-related validity. Findings suggest that while the scale effectively measures tense activation, further refinement is necessary to improve its comprehensiveness and precision.

Introduction

In the field of psychology, well-being is considered a fundamental component of human life and essential for achieving positive human health, the holistic state of optimal functioning (Ryff & Singer, 1998). Being a complex and multi-dimensional construct, well-being has been conceptualized in several ways. According to Diener (1984), subjective well-being encompasses three main components: life satisfaction, the presence of positive affect, and the absence of negative affect. Notably, in Diener's view, a recurring theme in many conceptualizations of well-being is affect. A general term, affect is defined as the underlying experience of feeling, emotion, attachment, or mood (Lazarus, 1991), making it a fundamental part of human experience. A specific instance of affect that includes both emotions and moods is called an affective state (Lühring et al., 2024). An example of an affective state is tense activation, one of the two activation systems distinguished by Thayer (1989). It is defined as a state of high activation due to stressors, accompanied by negative emotions, such as nervousness, feeling on edge, and tension. Physical effects such as muscle tension, an increased heartbeat, and increased stress hormones are also possible, preparing the body to handle the stressors. Consequently, a person enters a state of interconnected physiological and psychological tension, in reaction to perceived threats. Low tense activation consists of calmness and relaxation, with minimal tension or anxiety. In contrast, the other activation system, energetic activation, is defined as a state ranging from feelings of energy and vigour to tiredness and fatigue. This system highlights the dynamic nature of energy levels and their impact on an individual's ability to engage in activities and respond to environmental demands. Although Thayer's framework was initially contested and faced challenges, it was later validated by experimental data (Schimmack & Reisenzein, 2002). Consequently, these activation systems can be considered two distinct and independent dimensions. Besides being a verified independent activation system, tense activation, as an affective state, can be

interpreted by the circumplex model of Russell (1980), a model that categorizes affective states by valence and arousal. When applying this model, tense activation is characterized by negative valence and high arousal, while energetic activation is characterized by positive valence and high arousal. Also, although labelled differently, multiple other dimensions or affective states exhibit considerable overlap with tense activation. This is evident in Daniels' (2001) conceptualization of affective well-being, where he identifies five dimensions: anxiety, comfort, enthusiasm, depression, and vigour. The dimension of anxiety, similar to tense activation, includes emotions such as nervousness, feeling on edge, and tension. Another example is nervous tension, described by Spielberger (1972), which is characterized by feelings of tension, apprehension, and nervousness. These overlapping conceptualizations demonstrate that tense activation is a common and widely recognized affective state.

Considering the important role affect plays in well-being, it is crucial to effectively map and distinguish between affective states. The nuanced levels of affective states can provide valuable information to better understand their implications for well-being and potentially enhance it. In the case of tense activation, its components are predominantly associated with negative phenomena. For example, Ganster and Rosen (2013) found that stress-related feelings such as tension and anxiety are linked to negative health outcomes, while Kassel et al. (2003) identified associations between feelings of anxiety, tension, and nervousness, and substance use. As these variables may have serious effects, precise and valid measures of tense activation could be most useful.

Measuring Tense Activation

Although the dimension of “Anxiety” (Daniels, 2001) and the associated scale share conceptual similarities with tense activation, no measure has been specifically designed to assess tense activation. However, there are subscales of popular mood tests that cover this affective state. The “Tension-Anxiety subscale,” one of the six measured mood states in the

Profile off Mood States (POMS) (McNair et al., 1992), measures certain aspects of tense activation, with items like “I feel tense” and “I feel nervous”. Similarly, the State-Trait Anxiety Inventory (STAI) (Spielberger, 1983), includes a “State Anxiety scale” with items such as “I am tense” and “I am nervous”. Both scales focus on measuring short term affective states, reflecting on how one feels at a specific point in time. Additionally, these scores are applicable and informative in multiple contexts, underscoring the usefulness of measures of affective states

Written items are known to be effective measurement tools, but they come with certain disadvantages, such as the need for multiple measurement items. In Experience Sampling Methodology (ESM**)**, which is often used to assess moods, thoughts, symptoms, or behaviours that fluctuate over time (Ebner-Priemer et al., 2009), it is recommended that measurement durations be kept as short as possible to prevent survey fatigue among participants (Gabriel et al., 2019). Similarly, Ohly et al. (2010) state that diary studies should be designed to minimize participant burden to ensure the data collection process remains as efficient and manageable as possible. Furthermore, written survey items can cause cognitive burden when they include low-frequency words or complex sentence structures, leading to inaccurate responses (Lenzner et al., 2010).

A panacea to these issues may be the use of pictorial scales. Pictorial scales are typically a specific form of single-item measure, which can be efficient and reduce survey fatigue (Matthews et al., 2022). Compared to written items, pictorial scales are known to decrease cognitive load, as visual cues tend to simplify the process of understanding (Sauer, 2020). Summarizing the current disadvantages and gaps, a non-written measurement instrument for tense activation could be a highly useful addition. Therefore, to broaden the variety of measurement instruments for affective states, a single-item pictorial-scale that assesses tense activation will be developed and validated.

Pictorial Scales

A pictorial scale is a type of measurement tool that uses images or illustrations to help respondents indicate their feelings or perceptions, making it more engaging and easier to understand (Matthews et al., 2022). As pictorial scales have existed for a long time, multiple studies have examined their effects and psychometric characteristics. Papers reviewing research on pictorial scales (Baumgartner et al., 2019; Sauer et al., 2020) suggest that pictorial scales are more intuitively comprehensible, increase motivation, stimulate interest, and contain fewer errors in interpretation due to their language independency. Additionally, Bradley and Lang (1994) state that pictorial scale may be more appropriate in cross-cultural studies or when working with non-verbal populations. A common critique of pictorial scales is that these single-item measures fail to capture the whole construct, especially complex ones. However, this was debunked by Matthews et al., who showed that complex constructs can be reliably and validly assessed with single-item measures (2022). This can be explained by the fact that pictures are more intuitively comprehensible and can represent subjective states, making them less dependent on multiple items targeting certain aspects of the subjective state (Kunin, 1955; Broekens & Brinkman, 2013).

Typically, pictorial scales invoke metaphors to capture experiences. For example, the pain catastrophizing scale (PCS-C) (Crombez et al., 2003) uses a thermometer to reflect increasing levels of pain. The scale in this study is designed to resemble the stretching of a rubber band, reflecting the degree of tense activation that someone is experiencing, but also emphasizing the flexibility humans show in handling stressors. The rubber band is a universal object, primarily known for being under tension while still functioning, which makes it a logical and easily recognizable metaphor for tense activation. Additionally, this metaphor aligns well with the flexibility and stress-response dynamics described in psychological theories of emotional regulation (Gross 2002). To test the validity of this scale, two studies

will be conducted. First, an expert study will assess the content validity, including definitional alignment, comprehensiveness, and relevance. Second, the scale will be included in an ESM study to assess its convergent validity, discriminant validity and criterion-related validity.

Study 1. Expert Study On Content Validity

Haynes et al., (1995) defined content validity as the degree to which elements of an assessment instrument are relevant to and representative of the targeted construct for a particular assessment purpose, implying the constructs of relevance and comprehensiveness. To assess content validity, Hinkin and Tracey (1999) recommend gathering feedback on the degree of definitional correspondence, which is defined as the degree to which a scale's items align with the construct's definition. Feedback must be obtained from respondents who needed only one requirement "sufficient intellectual ability to rate the correspondence between items and definitions of various theoretical constructs, and the lack of any pertinent biases", and therefore were named "naïve judges". These types of respondents were considered useful as they are as less likely to be biased by their own knowledge and research. By using a Likert scale, this opinion on the degree of definitional correspondence is quantified, focusing on the mean score of the naïve judges. In this method, an odd number of Likert scale options is advised to achieve a more nuanced reflection of the naïve judges' opinions. This method was reviewed and modified in the paper by Colquitt et al. (2019), who transformed their formula to calculate an index, namely the *Hinkin Tracey Correspondence* (HTC), with a higher HTC indicating a more content-valid item. Colquitt et al. (2019) also provide guidelines for the interpretation of this statistic, defining a HTC of $< .59$ as lacking, between .60 and .83 as weak, between .84 and .86 as moderate, between .87 and .90 as strong, and .91 and above as very strong.

Another common way of establishing content validity is through the opinions of experts (Beck, 2020). This method is used by Lynn (1986) in the form of the Content Validity

Index (CVI), a quantitative measure that reflects the degree of content validity. This index is calculated separately for three constructs—relevance, comprehensiveness, and clarity—each measured using a Likert-style rating process. Unlike the HTC, an even number of Likert scale options is advised to encourage participants to take a position. The scores of this even number are then dichotomized into indicative or contraindicative of the measured construct. The proportion of indicative scores is then considered the CVI. According to the guidelines of Lynn (1986), an item is considered content valid if a CVI reaches the threshold of .83. Although not explicitly stated, it can be inferred that experts were preferred to rate the items due to their knowledge and experience with the subject.

To assess the content validity of this scale, parts of both methods will be used, leading to the measurement of comprehensiveness, relevance, and definitional correspondence. For simplicity, definitional correspondence was renamed to definitional alignment, as the question measuring this construct asked how well the scale aligns with the definition. Contrary to the Hinkin and Tracey method (1999), Lynn's (1986) approach of choosing experts over naïve judges will be adopted to invite a more critical view on the content validity of the rubber-band scale, potentially leading to more informative feedback. As the number of Likert-scale options must be chosen pre-emptively, literature on this number will be taken into consideration. DeMars and Erwin (2004) found that selecting the neutral response option **is** more likely with participants who are not familiar with the topic, do not have an opinion to report, or may not be interested in the topic. However, these factors are not likely to arise with experts. Considering other evidence supporting an odd number of Likert scale options (Johns, 2005; Kankaraš & Capecchi, 2024) and the potential informative value of adding an extra option, the Hinkin and Tracey method (1999) regarding the use of an odd number of Likert scale options will be applied in this study. To align with this choice, the method of Hinkin and Tracey (1999) will be the primary basis for assessing content validity. The focus of this study

will be on the calculated indices, their standard deviations, and the distribution of the experts' scores. In supplementary analyses, the method of Lynn (1986) will be applied to the data to achieve a multi-faceted view of the content validity. While comprehensiveness, relevance, and definitional alignment will be considered for calculating the indices, the construct of clarity will only be measured through optional qualitative comments, as it is considered a lower priority. This choice was made because the scale contains only one item, and the instructions are relatively short. Finally, the option will be provided to give supplementary feedback on other elements of the pictorial scale. This optional feedback could provide information for potential revisions to enhance the quality of the pictorial scale.

This multi-method approach is chosen because it provides the opportunity to view the content validity of the pictorial scale in a broader manner, instead of relying on a single construct to label the scale with a certain level of content validity. Since the indices are a simple transformation of scores of a continuous variable, the provided thresholds will not be considered strongly. Moreover, the indices can be considered as the degree of the measured construct, relative to the label of their answer options. Therefore, the exploratory and descriptive nature of this study leads to the following research questions:

R1: What is the definitional alignment (a), relevance (b), and comprehensiveness (c) of the rubber-band scale?

R2: What are the comments on the clarity of the instructions of the rubber-band scale?

R3: What are reoccurring themes of feedback for the rubber-band scale?

Method

Participants

To be considered eligible for joining the expert survey, participants needed to be the first author on a paper focused on work-related stress, strain, recovery, or burnout, published no earlier than May 2022. There were no further qualifications regarding age, sex, country of employment, or any other variables. To follow Lynn's (1986) method, supplemented by other literature on the minimum size of the sample of experts (Rubio et al., 2003; Shrotryia & Dhanda, 2019), a minimum of 6 experts was aimed for to conduct the expert survey. To ensure a minimum of 6 experts and, ideally, surpass this number, the threshold of invited participants was 150. Using the database "PsycINFO," participants were searched for using the keywords "work," "AND," "stress," "strain," "recovery," or "burnout." The authors that were shown first by this database were selected until the aim and threshold of 150 participants were reached. This led to inviting 150 participants, of which 32 eventually completed the survey. All participants finished the survey, and there were no reasons for exclusion of participants or deletion of data. The fast-track procedure was used in this study, whereby it was formally exempt from ethics board examination. No personal information of the participants was asked for or saved, as it was not needed. No reward was given for completing the survey.

Procedure

To establish the content validity of the rubber-band scale via expert ratings, a survey was designed using Qualtrics. After the research plan was approved by the Ethical Committee, the survey was activated, and the link for participation was sent out on October 6. At the start of the survey, the aim and scope of the study were presented to the experts. Additionally, the pictorials of the scale, the instructions, and the definition of tense arousal were presented. After presenting the definition of the construct, the instructions, and materials of the rubber-band scale, the experts were asked to rate the definitional alignment, relevance, and comprehensiveness of the rubber-band scale. For each rating, the experts were given the

opportunity to explain their answer. Finally, two open, non-mandatory questions were asked to assess the clarity of the instructions and to gather any additional general feedback. The survey was designed in such a manner that, when measuring each aspect, the rubber-band scale, the instructions, and the definition of tense arousal were available for reference, in case the participants needed them when answering the questions. Additionally, the definition of each construct was present when being measured. As expected, most responses were received within the first few days. Consequently, after two days without any new responses, the survey was deactivated on October 13.

Measures

Definitional alignment

The construct definitional alignment was measured with one question, namely: how well is the single-item pictorial scale aligned with the definition of tension presented above?". This question could be answered by a 5-point Likert-scale, consisting of "not aligned at all", "slightly aligned", "moderately aligned", "very aligned", and "completely aligned". Additionally, there was a non-mandatory follow-up question to explain the given answer. No definition of definitional alignment was presented.

Relevance

The construct relevance was measured with one question, accompanied by its definition (the extent to which the content is appropriate and useful for measuring tension), namely: "how relevant do you find the content for measuring the construct of tension?". This question could be answered by a 5-point Likert-scale, consisting of "not relevant at all", "slightly relevant", "moderately relevant", "very relevant", and "extremely relevant". Additionally, there was a non-mandatory follow-up question to explain the given answer.

Comprehensiveness

The construct comprehensiveness was measured with one question, accompanied by its definition (the extent to which the content covers all necessary aspects of the construct), namely: “how comprehensive do you find the content in covering the construct of tense arousal?”. This question could be answered by a 5-point Likert-scale, consisting of “not comprehensive at all”, “slightly comprehensive”, “moderately comprehensive”, “very comprehensive”, and “extremely comprehensive”. Additionally, there was a non-mandatory follow-up question to explain the given answer.

Data analysis

Hinkin and Tracey (1999) propose that judges rate how well scale items correspond to the construct’s definition using a 5-point Likert scale, whereby the resulting average rating provides a straightforward indication of definitional correspondence. Colquitt et al. (2019) modified this formula by dividing the total score by the total number of judges, resulting in the HTC (Hinkin Tracey Correspondence), which could maximally result in a 1 when all judges selected the maximum anchor for all scale items.

Lynn’s (1986) method will be applied in the supplementary analysis. Lynn (1986) proposed that experts should rate all items on their certain level of the measured construct (e.g., relevance), using an even-numbered Likert scale, where the scores are dichotomized and labelled as indicative of the measured construct. Dividing this number of indicative scores by the total number of experts leads to the item-level CVI (I-CVI), which in this study is labelled construct-level CVI (C-CVI), as there is only one item rated. When more than 5 experts rate the items, the C-CVI should be at least .83 (Lynn, 1986). This can be interpreted as meaning that 83% of the experts rate the measured construct with an indicative score, which means that the measured construct is highly present in the measured item. As this study uses an odd-numbered number of Likert-scale options, two variants of the method will be

used: one where the middle option is labelled as indicative of the measured construct and one where the middle option is labelled as contraindicative of the measured construct.

As there are possibly qualitative comments provided through the non-mandatory optional questions regarding written feedback, these will be analysed and categorized by content of the argument, based on the methodology of Brod et al. (2009). This article emphasizes the need for pattern recognition when analysing transcripts and focusing on finding recurring themes and categories in the data.

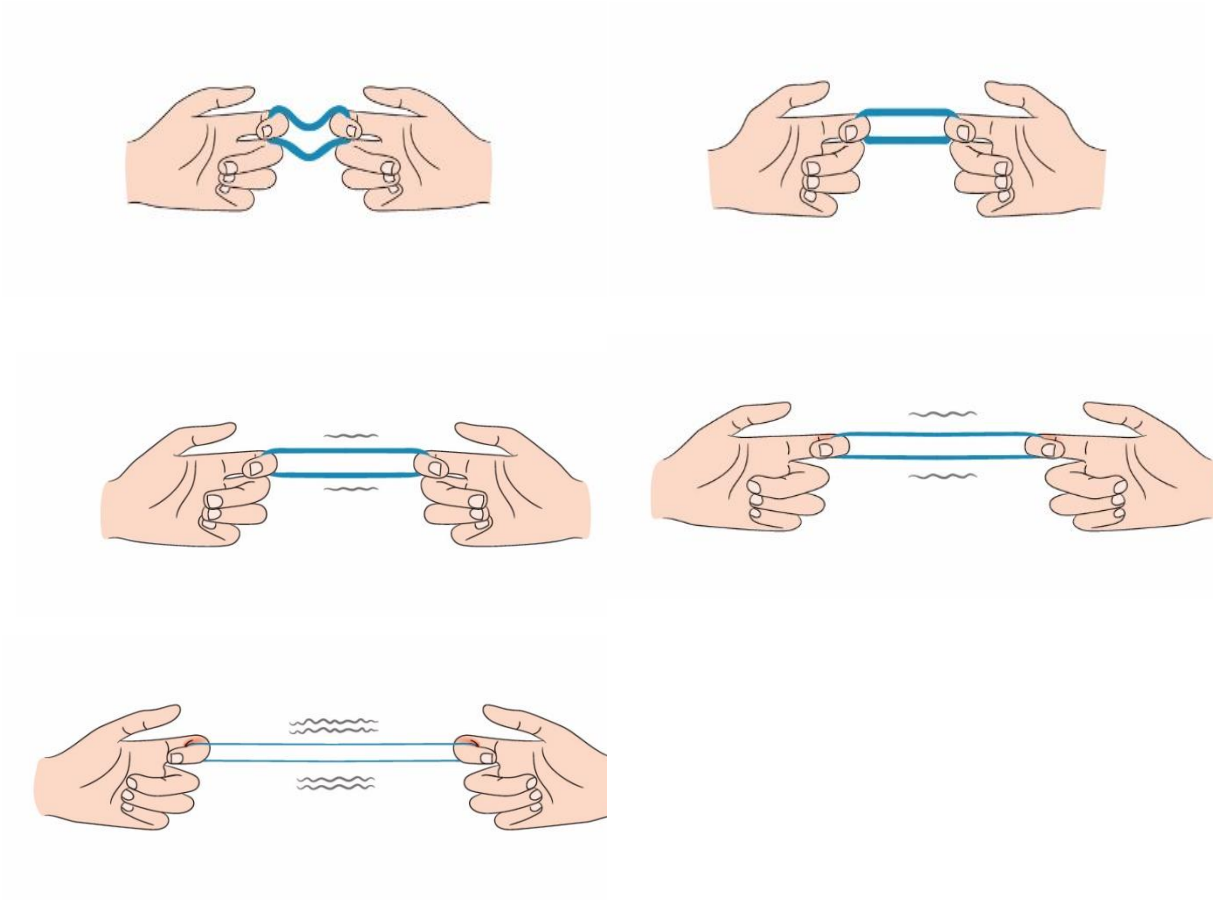
Materials

Rubber-band scale

The five pictorials of the rubber-band scale are depicted in Figure 1. The definition of tense arousal was provided, namely: “the degree to which a person is aroused by stressors, consisting of psychological and possibly physiological effects. The psychological effects are feelings of being on edge, anxiety and nervousness. The possible physiological effects are increased heart rate, muscle tension and sweating (Thayer, 1989)”. The instructions were: “the following pictures show a stretched rubber band to illustrate subjective states of tension ranging from feeling not tense at all to feeling extremely tense. Please rate which of the following symbols best describes how you feel right now”.

Figure 1

The five pictorials of the rubber-band scale



Note. The pictorials are arranged in an order where the degree of tense activation gradually increases, starting from the top-left, moving from left to right, and eventually reaching the bottom-left.

Results

Descriptives

The descriptives of the measured constructs definitional alignment, relevance and comprehensiveness are all presented in Table 1, which were obtained using SPSS. It appeared that the mean scores for all the measured constructs were above the midpoint (3.00) of the 5-point Likert scale, suggesting that the scale was generally viewed positively. While scoring the lowest average mean (M = 3.38), ratings of comprehensiveness also showed the most variability (SD = 1.04). The minimum score statistics elicit the fact that the lowest score of 1, referring to the complete absence of the measured construct, was not chosen a single time. The further distribution of the chosen answers is described in Table 2, which indicates that

option 4 was chosen most frequently for definitional alignment (p = .47) and relevance (p = .50), while option 3 was chosen most frequently for comprehensiveness (p = .38).

Table 1.

Descriptive statistics of definitional alignment, relevance and comprehensiveness

Construct	M	SD	Min	Max	N
Definitional alignment	3.66	0.90	2.00	5.00	32
Relevance	3.75	0.92	2.00	5.00	32
Comprehensiveness	3.38	1.04	2.00	5.00	32

Table 2.

Distribution of the C-CVI scores in proportions.

Construct proportion						
	Optional score	**1**	**2**	**3**	**4**	**5**
Definitional alignment		.00	.13	.25	.47	.16
Relevance		.00	.13	.19	.50	.19
Comprehensiveness		.00	.22	.38	.22	.19

Note. No perfect cumulation due to rounding. 1 = not at all. 2 = slightly. 3 = moderately. 4 = very. 5 = completely/extremely.

Hinkin and Tracey Correspondence

Using the paper by Colquitt et al. (2019) to calculate the HTC-statistic, the results are described in Table 4. Considering the guidelines of Colquitt et al. (2019) for interpreting the HTC-statistic, it appeared that all the calculated statistics of the rubber-band scale fell between the range of .60 to .83 (definitional alignment HTC = .73; relevance HTC = .75;

comprehensiveness HTC = .68), which is labelled as weak. Therefore, according to the approach by Colquitt et al. (2019), the rubber-band scale shows weak definitional alignment with tense activation, is weakly relevant in measuring tense activation, and is weakly comprehensive of the construct of tense activation, which answers research questions 1a, 1b, and 1c.

Table 3

HTC-statistics

Construct	**HTC-statistic**
Definitional alignment	.73
Relevance	.75
Comprehensiveness	.68

Construct-Level Content Validity Index

Using Lynn's (1986) method for calculating the C-CVI, we calculated the C-CVI in two variants, the results of which are presented in Table 3. Variant 1 consists of labelling only option 4 and 5 as indicative of the measured construct, and the middle option as contra-indicative. Variant 2 labels the middle option as indicative of the measured construct. This led to the output presented in Table 3. In variant 1, none of the measured constructs reached the proposed threshold by Lynn (1986) of .83 (definitional alignment C-CVI = .63, SD = .48; relevance C-CVI = .69, SD = .46; comprehensiveness C-CVI = .41, SD = .49). This implies that when approaching the calculation in a strict way, the rubber-band scale does not show sufficient content validity based on the measured constructs. In variant 2, the constructs of definitional alignment and relevance both reached and surpassed the threshold (definitional alignment C-CVI = .88, SD = .11; relevance C-CVI = .88, SD = .11), although

comprehensiveness did not (C-CVI = .78, SD = .41). This implies that when approaching the calculation in a lenient way, the rubber-band scale shows sufficient definitional alignment and relevance regarding content validity. These findings highlight how different variants of labelling the middle response option can significantly influence the perceived content validity of the rubber-band scale. It is noteworthy that comprehensiveness did not reach the threshold in any variant, although it came close in the second variant. Comprehensiveness also showed the highest standard deviations, reflecting the contrast of the experts' opinions on this matter.

Table 4

C-CVI statistics and standard deviations

Construct-CVI	**C-CVI variant 1**	**SD 1**	**C-CVI variant 2**	**SD 2**
Definitional alignment	.63	.48	.88	.11
Relevance	.69	.46	.88	.11
Comprehensiveness	.41	.49	.78	.41

Clarity and additional qualitative feedback

There were five comments that acknowledged the clarity of the instructions. An example of this is: "I don't have other feedback, the instructions are clear to me." On the contrary, there were two comments that directly or indirectly implied that the clarity of the instructions could be improved. An example of this is: "Maybe it should be mentioned in the instructions that an unpleasant state of tension is assessed and therefore described by the rubber band illustration. This is not intuitively clear as tension can be perceived positively as well (being awake, energetic, feeling strong or resilient)." Considering the five acknowledgments and all the experts who did not give suggestions for improvement of the

clarity of the instructions for the rubber-band scale, research question 2 can be answered as "sufficient."

Following the methodological advice of Brod et al. (2009), the additional qualitative feedback was categorized and led to the output described in Table 5. For the sake of redundancy and significance, only comments that had a clear point of critique or were a distinguishable acknowledgment were categorized and included in the table. This led to identifying ten categories of feedback. The three most frequently encountered categories were "Changing the Design" (N = 8), "Addition to Design" (N = 8), and "Not Capturing the (Full) Psychological Aspect" (N = 8). "Changing the design" included comments about altering the design of the pictorial scale, for example: "The redness on the fingers makes it look more focused on pain." "Addition to design" included comments about adding extra elements or details to the design of the pictorial scale, for example: "Sweat could maybe be displayed somehow, too." "Not capturing the (full) psychological aspect" included comments about the lack of measurement of the psychological aspect of tension, for example: "The aspect of anxiety is not very well captured in this picture." These top three most frequently mentioned categories can be used to answer research question 3. The category "Other" consists of statements that were in a category that did not reach more than one statement, for example: "The scale makes sense to me, while it is a relatively new way of measuring tension. I'm not so sure about the scale validity." The full coding of the transcripts is presented in Appendix A.

Table 5

Categories and frequencies of qualitative feedback

Category	**Frequency**

Changing the design	8
Addition to design	8
Not capturing the (full) psychological aspect	8
Acknowledging clarity of the instructions	5
Not capturing the (full) physiological aspect	5
Expanding the instructions	5
Too simplistic to capture the whole construct	4
Mentioning practically	4
Improving clarity of the instructions	2
Expanding the scale by items or options	2
Mentioning intuitiveness or straightforwardness	2
No additive value beyond Self-Assessment Manakin	2
Other	4

Discussion

The expert study on the content validity of the rubber-band scale gave mixed results. Descriptive statistics were fairly positive, as the options that indicate a strong presence (option 4; very) were chosen most often for the constructs of definitional alignment and relevance. For comprehensiveness, the most frequently chosen option was “moderate”, which is acceptable considering the breadth of the construct. Applying the Hinkin and Tracey’s (1999) method gave mediocre results, as this method led to labelling the presence of the measured constructs as weak. When interpreting the scores strictly using Lynn’s (1986)

method, the scale did not show sufficient presence in any constructs regarding content validity. Approaching the assessment in a more lenient way resulted in sufficient presence of the constructs “definitional alignment” and “relevance”. The threshold for comprehensiveness was not reached in either a lenient or strict manner using this method. The clarity of the instructions seemed to be sufficient. Aligning with the results of the construct of comprehensiveness, the most common feedback themes were addressing the lack of the physiological aspect of tense activation in the pictorial scale, advised changes for the design, and advised additions to the design. These results can be explained by considering the complex definitional nature of tense activation and the critical eye of experts. However, as pictorial scales depend on their intuitiveness, applying this scale in a different setting may shed light on other types of validities, which will be done in the second study.

Study 2. ESM Study for Convergent Validity, Discriminant Validity and Criterion-Related Validity

Incorporating the rubber-band scale into an ESM study offers a valuable opportunity to evaluate its psychometric properties by examining its relationships with other measurement items. Three types of validity will be assessed, namely convergent validity, discriminant validity, and criterion-related validity, through the interpretation of correlations. Correlation coefficients between .1 and .3 indicate a weak relationship, those between .3 and .5 indicate a moderate relationship, those between .5 and .7 indicate a strong relationship, and values between .7 and .9 indicate a very strong relationship (Schober et al., 2018).

Convergent validity, defined as the principle that tests measuring the same or similar constructs should be highly correlated (Convergent validity, 2011), will be assessed by comparing tense activation scores from the rubber-band scale (TARB) to a multi-item verbal tension scale and a single-item fatigue measure. Since this scale, whose content validity was established in the first study, measures tense activation characterized by feelings of tension, it

should logically correlate strongly with scores from verbal-written items assessing tension. This leads to the following hypothesis:

Hypothesis 1: Ratings of tense activation as captured with the rubber-band scale correlate strongly positively with tension as captured with a multi-item verbal scale for tension.

As Thayer (1989) distinguished two independent dimensions of activation, namely tense activation and energetic activation, he found differing effects for both types, such as the enhancement of mood versus the diminution of mood. Quinn et al., (2012) aligns with this distinction and furtherly underscores the positive valence of energetic activation and the negative valence of tense activation. Additionally, Russell's (1980) circumplex model classifies tense activation as negatively valenced and energetic activation as positively valenced. This distinction allows for the assessment of discriminant validity, defined as the extent to which a test does not correlate with measures of theoretically unrelated constructs (American Psychological Association, n.d.), by examining the correlation between TARB and energetic activation as measured by the battery scale (EABS) of Weigelt et al. (2022). Additionally, to further assess the discriminant validity of the rubber-band scale, the relationship between TARB and subjective vitality will be examined. Subjective vitality is defined as the subjective experience of possessing energy and aliveness (Ryan & Frederick, 1997). It has substantial conceptual overlap with energetic activation, making it relevant for assessing discriminant validity. Prior research has found correlations between measures of tension, subjective vitality, and energetic activation, ranging from -.37 to -.47 (Weigelt et al., 2022). Additionally, Miksza et al. (2019) found a significant negative relationship between stress, a construct similar to tense activation, and subjective vitality. Therefore, moderate negative links will be expected, which leads to the following hypothesis:

Hypothesis 2: Ratings of tense activation as captured with the rubber-band scale will correlate moderately negatively with (a) energetic activation as captured with the battery scale and (b) subjective vitality.

According to Hancock and Desmond (2001), fatigue is a psychological and physiological state that shares characteristics with the consequences of tense activation, such as common underlying mechanisms and impacts on performance. These include an increase in stress hormones and a decrease in cognitive abilities, such as memory retention. Other research (Kocalevent et al., 2011) found a significant positive association between stress and fatigue. Due to these similarities and the research of Kocalevent et al. (2011), fatigue is thought to correlate with TARB. However, following Thayer's (1989) dimensions of activation, fatigue is considered a low-activation state, while tense activation is a high-activation state. Therefore, only a moderate positive correlation is expected. This leads to the following hypothesis:

Hypothesis 3: Ratings of tense activation as captured with the rubber-band scale will correlate moderately positively with fatigue.

Finally, criterion-related validity, the degree of how well one measure predicts an outcome based on another measure (Vogt & Johnson, 2011), will be assessed by the correlation between the TARB and multiple other variables. These include workload, work goal progress, and relaxation and psychological detachment, all measured only in the evening. Higher workloads that exceed the ability to cope have been found to be related to increased psychological and physiological strain due to elevated levels of stress (Kosasih et al., 2024). Koudela-Hamila et al. (2022) identified a significant positive within-subject association between academic workload and academic stress. Considering the conceptual overlap of tense activation with stress, TARB and workload are expected to correlate positively. Exposure to stressors is known to be detrimental to performance and goal achievement due to the impact

of distractions, lower motivation, burnout, and cognitive overload (Sandi & Haller, 2015). Also, unfinished tasks at the end of the workweek lead to increased affective rumination, which consists of worrying and emotional distress (Syrek et al., 2017). This two-way interaction between stressors and work-goal progress—reflecting how much an individual is advancing toward achieving their work-related goals—highlights the incompatible functioning of these constructs. Therefore, work-goal progress will be expected to correlate negatively with TARB. Finally, relaxation and psychological detachment, two indicators of work recovery, are expected to correlate negatively with TARB. Increased work demands and stressors during the workday have been found to impede recovery experiences in the evening, such as psychological detachment and relaxation (Wendsche & Lohmann-Haislah, 2017). This study suggests that the inability to psychologically detach is hindered by stressors, as work-related thoughts continue even after work has ended. As a result, the lack of psychological detachment makes it difficult to achieve relaxation. Similarly, Sonnentag and Fritz (2007) found significant negative relationships between recovery experiences and job stressors. Based on these findings, the following hypotheses are proposed:

Hypothesis 4:Ratings of tense activation as captured with the rubber-band scale correlate moderately positively with workload.

Hypothesis 5: Ratings of tense activation as captured with the rubber-band scale correlate moderately negatively with work-goal progress.

Hypothesis 6: Ratings of tense activation as captured with the rubber-band scale correlate moderately negatively with a) relaxation and b) psychological detachment

Method

Participants

As 96 participants joined this study, the theoretical maximum of responses would be 5760 self-reports, considering the 15 days of measurement consisting of 4 measurement points. For the variables TARB, EABS, tension, and subjective vitality, 2961 self-reports, nested in 93 persons were available. Relative to the theorical maximum response, the response rate was 51.41%. As the variables fatigue, workload, work-goal progress, relaxation, and psychological detachment, were only measured once a day, 745 self-reports, nested in 85 persons were available. Relative to the theoretical maximum response of 1440 self-reports, the response rate was 51.74%.

Among 96 participants, 68 were female, 27 were male, and 1 identified otherwise. Ages ranged from 19 to 63 years (M = 33.10, SD = 9.07). Participants worked in the following industries: health/social services (36.46%), IT/communication (21.88%), education/culture (8.33%), finance/trade/logistics (7.29%), construction/trades (4.17%), automotive (1.04%), and other (20.83%).

Procedure

The ESM-study, in which the rubber-band scale was included, was a collaboration between the University of Groningen and the University of Hagen. Data were collected through a survey. Participants were asked to self-assess their levels of various constructs at four times a day: morning, noon, afternoon, and evening. This process continued for three consecutive weeks, excluding weekends. Instructions, as well as definitions of the scales and constructs to be measured, were provided at the beginning of the survey and remained accessible throughout its administration. Participants were recruited through the researchers' networks. The target sample size was 80 participants, with 800 self-reports per measurement point.

Measures

Rubber-band scale

The participants were asked to indicate which symbol best reflected their current state of tense activation. The choice had to be based on the five symbols, the instructions, and the definition of tense activation, as depicted in Figure 1. The instructions were: “the following pictures show a stretched rubber band to illustrate subjective states of tension ranging from feeling not tense at all to feeling extremely tense. Please rate which of the following symbols best describes how you feel right now”.

Tension

The construct of tension was measured with three items of the tension facet of the profile of mood states (POMS) (Mcnair et al., 1993). The scale has been adapted to German (Albani et al., 2005) and has been applied in experience sampling studies (Weigelt et al., 2022). These items focus on the current state of the participants. The three items item could be answered by a 5-point Likert-scale, reaching from “does not apply at all” to “fully applies”. The first item was: “I feel restless”. The second item was: “I feel uneasy". The third item was: “I feel tense”.

Energetic activation, subjective vitality and fatigue

The battery-scale is a validated measurement instrument for the construct of energetic activation (Weigelt et al. 2022). The participants were asked to indicate which symbol best reflected their current state of energetic activation. The choice was based on the seven symbols, the instructions, and the definition of energetic activation. The instructions were: “how one feels at the moment is often described in terms of the state of charge of a battery, ranging from “depleted” to “full of energy.” Please rate which of the following symbols best describes your current state”. The construct of subjective vitality was measured using three items from Ryan and Fredrick (1997), which were translated into German. These items focus

on the current state of the participants. The three items item could be answered by a 5-point Likert-scale, reaching from “does not apply at all” to “fully applies”. The first item was: “I feel alive and vital”. The second item was: “I feel energized”. The third item was: “I feel awake and alert”. The construct of fatigue was measured with three items of the POMS questionnaire (Albani et al., 2005). These items focus on the current state of the participants. The three items item could be answered by a 5-point Likert-scale, reaching from “does not apply at all” to “fully applies”. The first item was: “I feel worn out”. The second item was: “I feel weary”. The third item was: “I feel bushed”.

Workload and work-goal progress

The construct of workload was measured with one item of Spector and Jex (1998), which were translated to German. The item was: “today, there was a great deal to be done”. This item could be answered by a 5-point Likert-scale, reaching from “does not apply at all” to “fully applies”. The construct of work-goal progress was measured with one item of Koopman et al. (2016), which was translated to German. The item was: “"I made good progress with my work today”. This item could be answered by a 5-point Likert-scale, reaching from “does not apply at all” to “fully applies”.

Relaxation and psychological detachment

The construct of relaxation was measured with one item of Sonnentag and Fritz (2007), which was translated to German. The item was “I use the time to relax”. This item could be answered by a 5-point Likert-scale, reaching from “does not apply at all” to “fully applies. The construct of psychological detachment was measured with one item of Sonnentag and Fritz (2007), which was translated to German. The item was: “I distance myself from my work”. This item could be answered by a 5-point Likert-scale, reaching from “does not apply at all” to “fully applies.

Data analysis

The data is structured in four daily measurement points, in which certain variables are only measured once, namely in the evening. The calculation of the statistics will be carried out, based on the merging of the four data points. Multilevel confirmatory factor analyses will be used to analyse this ESM-data, in which the division of within-level and between-level correlations will be taken into account. The data is structured in Excel, will be transformed, and analysed using R version 4.4.2 (R Core Team, 2023), with the following packages: semTools (Jorgensen, 2021), Lavaan (Rosseel, 2012), and misty (Yanagida, 2024). Two models were specified to examine the hypotheses. The first model includes the variables that are measured on the four different measurement points. These are TARB, tension measured through the written items, EABS, and subjective vitality. The second model includes the variables that are measured only at the evening measurement point. These include fatigue, workload, work-goal progress, relaxation, psychological detachment, and the evening measurement of TARB. The hypotheses will be tested through the correlations between variables at the within-level. Correlations at the within-person level refer to the relationships between variables measured within the same individuals over time, whereas correlations at the between-person level refer to the relationships between variables measured across different individuals, capturing how one variable varies in relation to another across a group of people. Model fit indices, ICC's and multi-level composite reliabilities will be reported.

Preliminary Analysis

The descriptive statistics, multi-level reliabilities, and ICCs of model 1 are presented in Table 6. The ICC of the rubber-band scale appeared to be .40, indicating moderate reliability and implicating that 40% of the variability in measurements is due to true differences between subjects regarding levels of tense activation. The other ICCs ranged from .27 to .40, indicating fair to moderate reliability. All multi-item scales showed high

reliabilities, with no coefficient being lower than .89 at both within-person and between-person levels. The Chi-square test gave a significant result ($\chi^2 = 345.567$, $p < .01$) indicating that the model did not perfectly fit the data. Chi-square tests are known to be sensitive to larger sample sizes, which can explain the result. The Comparative Fit Index (CFI) and the Tucker-Lewis Index (TLI) both indicated that the data fits the model well (CFI = .98 and TLI = .97). The RMSEA and SRMR also indicated an acceptable and good model fit (RMSEA = .058, SRMR (within) = .021, and SRMR (between) = .033).

Convergent validity to a verbal measure of tension

Correlations are estimated from the MCFA and do not refer to observed values of tension. The correlation between TARB and tension captured through the written items appeared to be strongly positive (within-person level: $\psi = .66$, $p < .01$). Therefore, hypothesis 1 is confirmed.

Discriminant validity to measures of energetic activation and subjective vitality.

Correlations are estimated from the MCFA and do not refer to observed values of EABS and subjective vitality. The correlation between TARB and EABS appeared to be weakly negative (within-person level: $\psi = -.13$, $p = .03$). The correlation TARB and subjective vitality captured through the written items appeared to be insignificant (within-person level: $\psi = -.09$, $p = .13$). Therefore, hypothesis 2a is partially confirmed whether hypothesis 2b is rejected.

Table 6

Correlations, standard deviations, ICC's and multi-level composite reliabilities of model 1

Variable	1	2	3	4
1. Rubber-band scale	-	.80**	-.56**	-.46**
2. Tension	.66**	-	-.57**	-.42**

3. Battery scale	-.13*	-.22**	-	.81**
4. Subjective vitality	-.09	-.16**	.84**	-
Mean	2.09	2.23	4.57	2.94
SD within-person level	.79	.76	.08	.48
SD between-person level	.25	.25	.49	.40
ICC	.40	.40	.27	.28
MCA within-person level	-	.89	-	.91
MCA between-person level	-	.92	-	.93
MMO within-person level	-	.89	-	.91
MMO between-person level	-	.93	-	.93

Note. ICC = intra class correlation. MCA = multi-level Cronbach's Alpha. MMA = multi-level Macdonald's Omega. Correlations below the diagonal refer to the within-person level (N = 2961 self-reports). Correlations above the diagonal refer to the between-person level (N = 90 individuals). *. Correlations at the within-person level and correlations at the between-person level are significant at $p < .05$. **. Correlations at the within-person level and correlations at the between-person level are significant at $p < .01$.

Preliminary analysis

The descriptive statistics and multilevel reliabilities of model 2 are presented in Table 7. The ICCs ranged from .28 to .42, indicating fair to moderate reliability. The multi-item scale of fatigue showed high reliability, with no coefficient lower than .90 at both within-person and between-person levels. A Chi-square test yielded a non-significant result (χ^2 = 30.178, p = .07) suggesting that the model fits the data well. The Comparative Fit Index (CFI) and the Tucker-Lewis Index (TLI) both indicated excellent model fit (CFI = 1.00 and TLI = .99). The RMSEA and SRMR as well indicated a good model fit (RMSEA = .026, SRMR (within-person) = .014, and SRMR (between-person) = .033).

Convergent validity to a measure of fatigue

Correlations are estimated from the MCFA and do not refer to observed values of fatigue. The correlation between TARB and fatigue captured through the written items appeared to be weakly positive (within level: $\psi = .17$, $p = .01$). Therefore, hypothesis 3 is partially confirmed.

Criterion-related validity to measures of workload, work-goal progress, relaxation and psychological detachment

Correlations are estimated from the MCFA and do not refer to observed values of workload, work-goal progress, relaxation and psychological detachment. The correlation between TARB and workload captured through the written items appeared to be insignificant (within-person level: $\psi = .08$, $p = .07$). Therefore, hypothesis 4 is rejected. The correlation between TARB and work-goal progress captured through the written items appeared to be insignificant (within-person level: $\psi = .02$, $p = .64$). Therefore, hypothesis 5 is rejected. The correlation between TARB and relaxation captured through the written items appeared to be weakly negative (within-person level: $\psi = -.27$, $p < .01$). The correlation between TARB and psychological detachment captured through the written items appeared to be weakly negative (within-person level: $\psi = -.16$, $p < .01$). Therefore, hypothesis 6a is partially supported and hypothesis 6b is partially supported.

Table 7

Correlations, standard deviations, ICC's and multi-level composite reliabilities of model 2

Variable	**1**	**2**	**3**	**4**	**5**	**6**
1. Rubber-band scale	-	.45*	-.04	-.27	-.20	-.42*
2. Fatigue	.17**	-	.07	-.20	-.20	-.36*
3. Workload	.08	.17*	-	.55*	.29	.18

4. Work-goal progress	.02	-.03	.30*	-	.59*	.62*
5. Relaxation	-.27**	-.29*	-.09	.07	-	.78*
6. Psychological detachment	-.16**	-.28*	-.15*	.12*	.47*	-
Mean	1.85	2.86	3.18	3.61	4.25	3.59
SD within-person level	.55	.79	.96	.91	.97	.91
SD between-person level	.55	.50	.69	.42	.41	.19
ICC	.42	.30	.35	.29	.28	.39
MCA within-person level	-	.90	-	-	-	-
MCA between-person level	-	.92	-	-	-	-
MMO within-person level	-	.90	-	-	-	-

Note. ICC = intra class correlation. MCA = multi-level Cronbach's Alpha. MMA = multi-level Macdonald's Omega. Correlations below the diagonal refer to the within-person level (N = 745 self-reports). Correlations above the diagonal refer to the between-person level (N = 85 individuals). *. Correlations at the within-person level and correlations at the between-person level are significant at $p < .05$. **. Correlations at the within-person level and correlations at the between-person level are significant at $p < .01$.

Discussion

The study on convergent validity, discriminant validity and criterion-related validity, gave mixed, although primarily positive results. To establish convergent validity, the correlation between tense activation measured by the rubber-band scale (TARB) and written items of tension were expected to be strong. This correlation was found, thereby providing evidence for convergent validity. Discriminant validity was assessed by examining the correlation between TARB and EABS, and between TARB and subjective vitality. Moderate negative correlations were expected for both relationships. The correlation between TARB and EABS was weak but significant. The correlation between TARB and subjective vitality was insignificant. While the results for TARB and EABS support discriminant validity, those for TARB and subjective vitality do not. To expand the evidence on convergent validity, the

correlation between TARB and fatigue was assessed and expected to be moderately positive. This correlation appeared to be weak but significant. Despite being weaker than anticipated, it still provides evidence for convergent validity. Finally, to establish criterion-related validity, multiple correlations between TARB and criterion measures were assessed. The correlation between TARB and workload was expected to be moderately positive, but appeared to be insignificant. The correlation between TARB and work-goal progress was expected to be moderately negative, and appeared to be insignificant as well. These findings do not provide support for criterion-related validity. To further test criterion-related validity, the recovery experiences of relaxation and psychological detachment were expected to correlate moderately negatively with TARB. Both appeared to correlate weakly but significantly with TARB. These findings suggest that, although the relationships are weaker than expected, they still provide evidence for the criterion-related validity of TARB.

General discussion

This study was set out to validate a newly developed pictorial scale for measuring tense activation. The validation process aimed to ensure that the scale accurately and reliably measures tense activation across different contexts and populations. This was done through 2 studies; the first study was an expert study to assess the content validity of the rubber-band scale. In the second study, the rubber-band scale was applied in an ESM study, to assess the convergent validity, discriminant validity, and criterion-related validity.

Theoretical implications

The first study gave mixed results for content validity. Based on the index of Hinkin and Tracey's (1999) method and the strictly measured index of Lynn's (1986) method, the rubber-band scale showed weak content validity. However, it has to be noted that these indices are merely a transformation of the chosen Likert-scale options. The distribution of the

scores therefore have to be taken relative to label of these scores. For the constructs of definitional alignment and relevance, the option “very”, for degree of presence of the measured construct, was by far the most often chosen answer. Also, the label of the fifth option was “completely”, which according to the methods would indicate perfect content validity, wasn’t the least chosen option for the constructs of definitional alignment and relevance (“not at all” and then “slightly” were). The effect of the labelling was also reflected in the C-CVI’s of the lenient interpretation of Lynn’s method, where both definitional alignment and relevance reached and surpassed the thresholds for content validity. Another note is that subject experts were used to rate the content validity, unlike naïve judges used in the Hinkin and Tracey (1999) method. Although this is debatable, experts are generally considered to be more critical in rating scales, as they have more knowledge on the subject and a keener eye for details. More knowledge on the subject could help identify certain aspects that were missing or had a lower quality, resulting in a more critical evaluation of the scale and therefore lead to lower indices. Comprehensiveness appeared to be the least present construct in the rubber-band scale. This issue was highlighted by the qualitative feedback, which signified the scale's inability to fully capture the physiological aspect. A possible explanation for this is the breadth of the construct and the fact that the pictorial scale has only one item, which makes it difficult to completely cover all aspects of the measured affective state. This aligns with common critiques of pictorial scales, namely its difficulty with comprehensively capturing complex constructs (Matthews et al., 2022). However, research indicated that pictorial scales are easier understanded through intuition and therefore may not be dependent on multiple items (Kunin, 1955; Broekens & Brinkman, 2013). Thus, while the single-item format of the pictorial scale may limit its ability to capture the construct's full complexity, its intuitive design could enhance user comprehension and ease of use. This first

study highlights the challenge of measuring a broad affective state with a single-item pictorial scale.

The second study on the convergent validity, discriminant validity, and criterion-related validity gave mixed results. As the scores of TARB correlated strongly with scores on written items of tension, convergent validity appears to be presents, which indicates that the rubber-band scale does measure it's intended construct. Also, the scores of TARB and fatigue correlated weakly but significant, which adds support for the evidence of convergent validity. The correlation between TARB and written items of tension, suggest that the rubber-band measures an affective state of which tension is a significant part, implying the effectiveness of the rubber-band scale in measuring tense activation. The correlation between TARB and fatigue reflects that the presence of indicators of a low energetic activation are associated with higher levels of tense activation, emphasizing the distinction between the two dimensions of Thayer (1989). The weak, although significant negative correlation of TARB with EABS, shows the presence of discriminant validity in the rubber-band scale. This indicates that the rubber-band scale effectively measures a construct distinct of energetic activation. Considering the findings on convergent validity, this appears to be tense activation. This finding aligns with the separation of the two dimensions by Thayer (1989) and the experimental verification of Schimmack and Reisenzein, (2002). The lack of association between the rubber-band scale scores and written items of subjective vitality do not provide evidence for discriminant validity. The absence of a significant negative correlation at the within-level indicates that these variables do not fluctuate similarly for individuals. However, the strong negative association at the between-person level ($\psi = -.56$, $p < .01$) signifies that, across the entire sample, these variables tend to vary inversely. A possible explanation for this differing result is the response format of the items. Due to the fact that the pictorial scales have similar visual formats, it possibly made it more inviting to coherently answer these

items. The discrepancy between the response formats of TARB and subjective vitality may have led to varied interpretations of the answer options, resulting in less consistent responses and, consequently, influencing the observed differences in correlations at both the within- and between-person levels. Regarding criterion-related validity, no associations were found between TARB and scores of workload and work-goal progress. These findings suggest that the rubber-band scale may not be effective in predicting or reflecting these specific work-related outcomes. A theoretical explanation for these findings, may be the fact that less desired work-related outcomes do not necessarily evoke emotions related to tense-activation. Rather, these could be viewed as a challenge and energize people. This aligns with the Challenge-Hindrance Stressors Framework (Cavanaugh, 2000), which proposes that if stressors are perceived as opportunities, they are associated with positive outcomes such as increased motivation, job satisfaction, and performance. The correlations between TARB and recovery experiences were smaller than expected, but significant. Still, this suggests that the rubber-band scale measures a construct that is significantly negatively associated with constructs that measure how well someone is recovering from work-activities. Based on the literature regarding how recovery experiences can be hindered (Sonnentag & Fritz, 2015; Sonnentag & Fritz, 2007), it is likely that the construct negatively associated with recovery is conceptually similar to tense activation. However, the associations were weaker than expected, which could potentially be explained by the timing of when the recovery experiences were measured. As acquired data points were measured only once a day, and varied from times as 19:00 to 23:00, this could have been a source of erroneous variation. In the course of the evening, multiple variables could have enhanced the recovery experiences, and cancel out tense activation, for example by social interactions (Sonnentag & Fritz, 2015).

Practical implications

While there is room for improvement, the rubber-band scale is useful in many different contexts when its limitations are taken into account. First of all, because it has important implications for both employers and employees, tense activation is a highly relevant construct in the workplace. There are several benefits of using a low-cost instrument to measure tense activation. For instance, it can assist in identifying early indicators of employee burnout, enabling customized care to avoid more serious mental health problems. It can also identify tasks or situations that trigger high levels of tense activation, offering insights for enhancing job satisfaction and workplace attitudes. By addressing these factors, this approach not only safeguards employees' well-being but also improves productivity and reduces absenteeism, ultimately benefiting the entire company. Secondly, the rubber-band scale can be used in clinical settings to track tense activation levels in patients with conditions like anxiety disorders, depression, and PTSD throughout their treatment. By monitoring these levels, clinicians can gain insights into the patient's psychological state and the effectiveness of the interventions. This continuous tracking allows for timely adjustments to treatment plans, ensuring that these approaches remain responsive to the patient's needs. Thirdly, the rubber-band scale can serve as both a primary and complementary tool in research settings. As a primary tool, it can be used to track levels of tense activation in studies specifically focused on this construct, providing data on how tense activation varies across different conditions and populations. As a complementary tool, it can be used alongside other measurement instruments to compare and validate their effectiveness. By assessing the validity of other tools against the rubber-band scale, researchers can ensure that their methods are accurately capturing the intended constructs. Fourth, a more general view on the practicality of this scale highlights its simplicity and universality. Since the scale consists of only one pictorial item, it is an economical measure with a low cognitive load on participants, making it practical and easy to use in various settings. Additionally, its language-independent

nature and intuitive pictorial design make it easy to understand, making it applicable across different cultures and age groups.

Strengths and limitations

The fact that two studies were conducted on different types of validity provided a better understanding and more comprehensive view on the psychometric characteristics of the rubber-band scale. Additionally, differentiating between experts and non-experts offered more in-depth insights. The sample of experts significantly exceeded the minimum required, providing a rich source of information on which implications of content validity could be based. This was also true for the second study, as both models had high response rates, with a sample size of 96 individuals providing self-reports. Interestingly, the large sample of experts revealed a discrepancy in opinions. While many opinions fell under the same umbrella, there were also multiple opinions that contradicted the assessments of other experts. An example of this is the coverage of the physical or psychological aspects, where both compliments and criticisms were given on both sides. This emphasizes the differing opinions of various experts, which must be taken into account when basing the content validity on these assessments.

A potential limitation of this study is that only tension, as a component of tense activation, was assessed for convergent validity in the second study. Since tense activation is a broad affective state, including emotions such as nervousness and anxiety would provide a more comprehensive assessment of the convergent validity of the rubber-band scale. Additionally, the quantity of the qualitative feedback could have been more if the optional feedback question had been made mandatory. This would have ensured that all participants provided their insights, leading to a richer dataset. Furthermore, deviating from both the Hinkin and Tracey (1999) method and Lynn's (1986) method may have compromised the validity of the data. Although the changes were minor, modifying certain aspects of this method and deviating from the proposed guidelines, such as using experts instead of naïve

judges, could impact the interpretation of the resulting statistics. These deviations are probably not substantial, but should be kept in mind when interpreting the results of this study.

Directions for future research

Since this was the initial validation of the rubber-band scale, there is significant room for development. To improve its psychometric characteristics, user-friendliness, and other qualities, time must be dedicated to refining the scale. The collected expert feedback could be valuable in this process. Given the numerous suggestions for changes or additions to this pictorial scale, there are multiple areas which can be reviewed. As especially the physiological aspect of the scale seemed to be incomplete, different variants of this scale could be designed with changed or added elements that emphasize this aspect more. Although many psychometric characteristics have been evaluated, further research could focus on the test-retest reliability and factorial structure of the scale. Test-retest reliability examines the consistency of results over time, which can assess the stability of this scale. Investigating the factorial structure could help to understand other possible underlying dimensions the scale measures. These analyses could confirm its validity and accuracy in measuring the intended constructs.

A next step in research could involve applying the scale in longitudinal studies. In these studies, researchers could observe how the scale performs across extended periods, providing insights into its consistency and reliability. Also, these studies could reveal how the scale interacts with different external influences and situations that participants encounter over time. This approach would help determine whether the scale can accurately predict outcomes and maintain its validity in different contexts and conditions. Additionally, this scale could be applied across different cultures and age groups to evaluate its performance in diverse samples. By testing the scale in various cultural contexts and among different age

groups, researchers can assess its generalizability and ensure that it accurately measures the intended constructs regardless of the sample's background. This approach would help determine the scale's robustness and adaptability, providing insights in how it functions across different settings.

Lastly, future research could explore how well the rubber-band scale converges with other dimensions of tense activation, beyond tension. This would provide deeper insights into the scale's ability to capture the full spectrum of the affective state of tense activation. Besides the psychological aspect, the scale’s effectiveness in measuring the physiological manifestations of tense activation could also be tested. For example, a study could have participants complete the rubber-band scale while measuring physiological responses like heart rate or muscle tension to assess alignment.

Conclusion

The aim of these studies was to develop and validate a newly designed pictorial scale for tense activation, namely the rubber-band scale. In summary, the findings clarify the different types of validity associated with measuring tense activation using the rubber-band scale. The expert study indicated a mild degree of content validity; however, when considering the distribution of the experts' scores relative to the labels, one could argue that multiple constructs of content validity are strongly represented in the rubber-band scale. The inclusion of the rubber-band scale in an ESM study showed promising results, where findings implied evidence for convergent validity, discriminant validity, and criterion-related validity. Considering the breadth and complexity of the construct, this single-item measure can serve as a useful instrument in the assessment of tense activation, in a quick and economical way.

References

Albani, C., Blaser, G., Geyer, M., Schmutzer, G., Brähler, E., Bailer, H., & Grulke, N. (2005). Überprüfung der Gütekriterien der deutschen Kurzform des Fragebogens "Profile of Mood States" (POMS) in einer repräsentativen Bevölkerungsstichprobe [The German short version of "Profile of Mood States" (POMS): Psychometric evaluation in a representative sample]. *Psychotherapie, Psychosomatik, Medizinische Psychologie, 55*(7), 324–330. https://doi.org/10.1055/s-2004-834727

American Psychological Association. (2023). *Stress effects on the body*. Retrieved January 31, 2025, from https://www.apa.org/topics/stress/body

American Psychological Association, APA Task Force on Psychological Assessment and Evaluation Guidelines. (2020). *APA guidelines for psychological assessment and evaluation*. Retrieved January 5, 2025, from https://www.apa.org/about/policy/guidelines-psychological-assessment-evaluation.pdf.

American Psychological Association. (n.d.). Discriminant validity. In *APA dictionary of psychology*. Retrieved December 17, 2024, from https://dictionary.apa.org/discriminant-validity.

Baumgartner, J., Sonderegger, A., & Sauer, J. (2019). No need to read: Developing a pictorial single-item scale for measuring perceived usability. *International Journal of Human-Computer Studies, 122*, 78-89. https://doi.org/10.1016/j.ijhcs.2018.08.004

Beck, K. (2020). Ensuring content validity of psychological and educational tests – The role of experts. *Frontline Learning Research, 8*(6), 1–37. https://doi.org/10.14786/flr.v8i6.517

Bradley, M. M., & Lang, P. J. (1994). Measuring emotion: The Self-Assessment Manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry, 25*(1), 49–59. https://doi.org/10.1016/0005-7916(94)90063-9

Brod, M., Tesler, L. E., & Christensen, T. L. (2009). Qualitative research and content validity: Developing best practices based on science and experience. *Quality of Life Research, 18*(9), 1263-1278. https://doi.org/10.1007/s11136-009-9540-9

Broekens, J., & Brinkman, W.-P. (2013). AffectButton: A method for reliable and valid affective self-report. *International Journal of Human-Computer Studies, 71*(6), 641-667. https://doi.org/10.1016/j.ijhcs.2013.02.003

Cavanaugh, M. A., Boswell, W. R., Roehling, M. V., & Boudreau, J. W. (2000). An empirical examination of self-reported work stress among U.S. managers. *Journal of Applied Psychology, 85*(1), 65-74. https://doi.org/10.1037/0021-9010.85.1.65

Colquitt, J. A., Sabey, T. B., Rodell, J. B., & Hill, E. T. (2019). Content validation guidelines: Evaluation criteria for definitional correspondence and definitional distinctiveness. *Journal of Applied Psychology, 104*(10), 1243–1265. https://doi.org/10.1037/apl0000406

Convergent validity. (2011). In J. S. Kreutzer, J. DeLuca, & B. Caplan (Eds.), *Encyclopedia of Clinical Neuropsychology*. Springer. https://doi.org/10.1007/978-94-007-0753-5_573

Crombez, G., Bijttebier, P., Eccleston, C., Mascagni, T., Mertens, G., Goubert, L., & Verstraeten, K. (2003). The child version of the pain catastrophizing scale (PCS-C): A preliminary validation. *Pain, 104*(3), 639-646. https://doi.org/10.1016/S0304-3959(03)00121-0

Daniels, K. (2000). Measures of five aspects of affective well-being at work. *Human Relations, 53*(2), 275-294. https://doi.org/10.1177/0018726700532004

DeMars, C. E., & Erwin, T. D. (2004). Scoring "neutral or unsure" on an identity development instrument for higher education. *Research in Higher Education, 45*(1), 83–95. https://doi.org/10.1023/B:RIHE.0000010049.19494.a2

Diener, E. (1984). Subjective well-being. *Psychological Bulletin, 95*(3), 542-575. https://doi.org/10.1037/0033-2909.95.3.542

Ebner-Priemer, U. W., Eid, M., Kleindienst, N., Stabenow, S., & Trull, T. J. (2009). Analytic strategies for understanding affective (in)stability and other dynamic processes in psychopathology. *Journal of Abnormal Psychology, 118*(1), 195-202. https://doi.org/10.1037/a0014868

Gabriel, A. S., Podsakoff, N. P., Beal, D. J., Scott, B. A., Sonnentag, S., Trougakos, J. P., & Butts, M. M. (2019). Experience sampling methods: A discussion of critical trends and considerations for scholarly advancement. *Organizational Research Methods, 22*(4), 969–1006. https://doi.org/10.1177/1094428118802626

Ganster, D. C., & Rosen, C. C. (2013). Work stress and employee health: A multidisciplinary review. *Journal of Management, 39*(5), 1085-1122. https://doi.org/10.1177/0149206313475815

Gross, J. J. (2002). Emotion regulation: Affective, cognitive, and social consequences. *Psychophysiology, 39(3)*, 281-291. https://doi.org/10.1017/S0048577201393198

Hancock, P. A., & Desmond, P. A. (2001). Stress and fatigue in human performance. *Human Factors, 43*(2), 189-206. https://doi.org/10.1518/001872001775387271

Haynes, S. N., Richard, D. C. S., & Kubany, E. S. (1995). Content validity in psychological assessment: A functional approach to concepts and methods. *Psychological Assessment, 7*(3), 238–247. https://doi.org/10.1037/1040-3590.7.3.238

Hinkin, T. R., & Tracey, J. B. (1999). An analysis of variance approach to content validation. *Organizational Research Methods, 2*(2), 175–186. http://dx.doi.org/10.1177/109442819922004

Johns, R. A. (2005). One size doesn't fit all: Selecting response scales for BES attitude items. *Journal of Elections, Public Opinion and Parties, 15*(2), 237-264. https://doi.org/10.1080/13689880500178849

Jorgensen, T. D., Pornprasertmanit, S., Schoemann, A. M., & Rosseel, Y. (2021). Useful tools for structural equation modeling [R package semTools version 0.5-4] [Computer software]. Comprehensive R Archive Network (CRAN). https://CRAN.R-project.org/package=semTools

Kankaraš, M., & Capecchi, S. (2024). Neither agree nor disagree: Use and misuse of the neutral response category in Likert-type scales. *Journal of Survey Research, 12*(3), 45-67. https://doi.org/10.1007/s40300-024-00276-5

Kassel, J. D., Stroud, L. R., & Paronis, C. A. (2003). Smoking, stress, and negative affect: Correlation, causation, and context across stages of smoking. *Psychological Bulletin, 129*(2), 270-304. https://doi.org/10.1037/0033-2909.129.2.270

Koopman, J., Lanaj, K., & Scott, B. A. (2016). Integrating the bright and dark sides of OCB: A daily investigation of the benefits and costs of helping others. *Academy of Management Journal, 59*(2), 414–435. https://doi.org/10.5465/amj.2014.0262

Kocalevent, R. D., Hinz, A., Brähler, E., & Klapp, B. F. (2011). Determinants of fatigue and stress. *Journal of Psychosomatic Research, 71*(3), 159-164. https://doi.org/10.1016/j.jpsychores.2011.01.009

Kosasih, L., Judijanto, L., Pasagi, Y., Dewi, I. C., & Hartono, S. (2024). The moderating effect of leadership on the relationship between workload and work stress: Empirical evidence from a public hospital. *Journal of Logistics, Informatics and Service Science, 11*(8), 1-12. https://doi.org/10.33168/JLISS.2024.0801

Koudela-Hamila, S., Santangelo, P. S., Ebner-Priemer, U. W., & Schlotz, W. (2022). Under which circumstances does academic workload lead to stress? Explaining intraindividual differences by using the cortisol-awakening response as a moderator. *Journal of Psychophysiology, 36*(3), 188–197. https://doi-org.proxy-ub.rug.nl/10.1027/0269-8803/a000293

Kunin, T. (1955). The construction of a new type of attitude measure. *Personnel Psychology, 8*, 65–77. https://doi.org/10.1111/j.1744-6570.1955.tb01189.x

Lazarus, R. S. (1991). *Emotion and adaptation*. Oxford University Press.

Lenzner, T., Kaczmirek, L., & Lenzner, A. (2010). Cognitive burden of survey questions and response times: A psycholinguistic experiment. *Applied Cognitive Psychology, 24*(7), 1003–1020. https://doi.org/10.1002/acp.1602

Lovibond, P. F., & Lovibond, S. H. (1995). The structure of negative emotional states: Comparison of the Depression Anxiety Stress Scales (DASS) with the Beck Depression and Anxiety Inventories. *Behavior Research and Therapy*, 33, 335–343. https://doi-org.proxy-ub.rug.nl/10.1016/0005-7967(94)00075-U

Lühring, J., Shetty, A., Koschmieder, C., Garcia, D., Waldherr, A., & Metzler, H. (2024). Emotions in misinformation studies: Distinguishing affective state from emotional response and misinformation recognition from acceptance. Journal of Experimental Psychology: General, 153(2), 123-145. https://doi.org/10.1037/xge0001234

Lynn, M. R. (1986). Determination and quantification of content validity. *Nursing Research*, *35*(6), 382–385. https://doi.org/10.1097/00006199-198611000-00017

Matthews, R. A., Pineault, L., & Hong, Y.-H. (2022). Normalizing the use of single-item measures: Validation of the single-item compendium for organizational psychology. *Journal of Business and Psychology*, *37*(3), 639-673. https://doi.org/10.1007/s10869-022-09813-3

McNair, D. M., Lorr, M., & Droppleman, L. F. (1971). *Manual for the Profile of Mood States (POMS)*. San Diego, CA: Educational and Industrial Testing Service

Miksza, P., Evans, P., & McPherson, G. (2019). Wellness among university-level music students: A study of the predictors of subjective vitality. *Psychology of Music, 47*(4), 407-423. https://doi.org/10.1177/0305735619854530

Ohly, S., Sonnentag, S., Niessen, C., & Zapf, D. (2010). Diary studies in organizational research: An introduction and some practical recommendations. *Journal of Personnel Psychology*, *9*(2), 79–93. https://doi.org/10.1027/1866-5888/a000009

Quinn, R. W., Spreitzer, G. M., & Lam, C. F. (2012). Building a sustainable model of human energy in organizations: Exploring the critical role of resources. *The Academy of Management Annals*, *6*(1), 337–396. https://doi-org.proxy-ub.rug.nl/10.1080/19416520.2012.676762

R Core Team. (2023). *R: A language and environment for statistical computing* (Version 4.4.2) [Software]. R Foundation for Statistical Computing. https://www.R-project.org/

Rosseel, Y. (2012). Lavaan: An R package for structural equation modeling. *Journal of Statistical Software, 48*(1), 1–36. https://doi.org/10.18637/jss.v048.i02

Rubio, D. M., Berg-Weger, M., Tebb, S. S., Lee, E. S., & Rauch, S. (2003). Objectifying content validity: Conducting a content validity study in social work research. *Social Work Research, 27*(2), 94-104. https://doi.org/10.1093/swr/27.2.94

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology, 39*(6), 1161-1178. https://doi.org/10.1037/h0077714

Ryan, R. M., & Frederick, C. M. (1997). The energy behind human flourishing: Theory and research on subjective vitality. *Journal of Personality, 65*(3), 529-565. https://doi.org/10.1111/j.1467-6494.1997.tb00326.x

Ryff, C. D., & Singer, B. H. (1998). The contours of positive human health. *Psychological Inquiry, 9*(1), 1–28. https://doi.org/10.1207/s15327965pli0901_1

Sandi, C., & Haller, J. (2015). Stress and the social brain: Behavioural effects and neurobiological mechanisms. *Nature Reviews Neuroscience*, *16*(5), 290-304. https://doi.org/10.1038/nrn3918

Sauer, J., Baumgartner, J., Frei, N., & Sonderegger, A. (2020). Pictorial scales in research and practice: A review. *European Psychologist, 26*(3), 1-19. https://doi.org/10.1027/1016-9040/a000405

Schimmack, U., & Reisenzein, R. (2002). Experiencing activation: Energetic arousal and tense arousal are not mixtures of valence and activation. *Emotion, 2*(4), 412-417. https://doi.org/10.1037/1528-3542.2.4.412

Schober, P., Boer, C., & Schwarte, L. A. (2018). Correlation coefficients: Appropriate use and interpretation. *Anesthesia & Analgesia, 126*(5), 1763–1768. https://doi.org/10.1213/ANE.0000000000002864

Shrotryia, V. K., & Dhanda, U. (2019). Content validity of assessment instrument for employee engagement. *SAGE Open, 9*(1), 2158244018821751. https://doi.org/10.1177/2158244018821751

Sonnentag, S., & Fritz, C. (2007). The Recovery Experience Questionnaire: Development and validation of a measure for assessing recuperation and unwinding from work. *Journal of Occupational Health Psychology, 12*(3), 204–221. https://doi.org/10.1037/1076-8998.12.3.204

Sonnentag, S., & Fritz, C. (2015). Recovery from job stress: The stressor-detachment model as an integrative framework. *Journal of Organizational Behavior, 36*(S1), S72–S103. https://doi.org/10.1002/job.1924

Spector, P. E., & Jex, S. M. (1998). Development of four self-report measures of job stressors and strain: Interpersonal Conflict at Work Scale, Organizational Constraints Scale, Quantitative Workload Inventory, and Physical Symptoms Inventory. *Journal of Occupational Health Psychology, 3*(4), 356-367. https://doi.org/10.1037/1076-8998.3.4.356

Spielberger, C. D. (1972). The effects of anxiety on experimental task performance. In L. Berkowitz (Ed.), *Advances in Experimental Social Psychology* (Vol. 6, pp. 1-40). Academic Press.

Spielberger, C. D., Gorsuch, R. L., Lushene, R., Vagg, P. R., & Jacobs, G. A. (1983). *Manual for the State-Trait Anxiety Inventory (Form Y)*. Palo Alto, CA: Consulting Psychologists Press.

Syrek, C. J., Weigelt, O., Peifer, C., & Antoni, C. H. (2017). Zeigarnik's sleepless nights: How unfinished tasks at the end of the week impair employee sleep on the weekend through rumination. *Journal of Occupational Health Psychology, 22*(2), 225–238. https://doi.org/10.1037/ocp0000031

Thayer, R. E. (1989). *The biopsychology of mood and arousal*. Oxford University Press.

Vogt, W. P., & Johnson, R. B. (2011). *Dictionary of Statistics & Methodology: A Nontechnical Guide for the Social Sciences*. SAGE.

Weigelt, O., Syrek, C. J., Kühnel, J., & Vahle-Hinz, T. (2022). Time to recharge batteries: Development and validation of a pictorial scale of human energy. *European Journal of Work and Organizational Psychology, 31*(5), 694–706. https://doi.org/10.1080/1359432X.2022.2050218

Wendsche, J., & Lohmann-Haislah, A. (2017). A meta-analysis on antecedents and outcomes of detachment from work. *Frontiers in Psychology*, *7*, 2072. https://doi.org/10.3389/fpsyg.2016.02072

Yanagida, T. (2024). *misty: Miscellaneous functions for structural equation modeling* (Version 0.6.8) [R package]. Comprehensive R Archive Network (CRAN). https://CRAN.R-project.org/package=misty

Appendix A. Coding scheme of qualitative feedback

Category of question regarding additional feedback	Statement	Coding category
Definitional alignment	"The arousal and anxiety part does not really seem to be reflected in the item. I'm not sure if the rubber band is the best way to display these feelings."	Not capturing the (full) psychological aspect/ changing the design
Definitional alignment	"While I see the tension part in terms of being very tense (angespannt), I was not sure about the anxiety and nervousness described in the definition. I do not think of anxiety when looking at the pictures."	Not capturing the (full) psychological aspect
Definitional alignment	"Anxiety and nervousness can be perceived very differently. Not all individuals would use the analogy of a stretched rubber band to illustrate this state, even though the definition is "tension". "Tension" describes one subjective state of experiencing the consequences of being exposed to stressors."	Other
Definitional alignment	"The redness on the fingers makes it look more focused on pain."	Changing the design
Definitional alignment	"Because the pictorial scale uses an easy-to-understand metaphor for "tension""	Mentioning intuitiveness or straightforwardness/ mentioning practicality
Definitional alignment	"The single-item pictorial scale aligns well with the definition of psychological tension in terms of representing subjective states of arousal. It effectively captures increasing psychological stress (nervousness and anxiety). However, it may be less effective in representing the physiological aspects of tension. While the rubber band	Not capturing the (full) physiological aspect/ mentioning intuitiveness or straightforwardness

	metaphor indirectly represents physiological effects (e.g., muscle tension), it doesn't explicitly evoke heart rate or sweating, which are mentioned in the definition. Overall, the rubber band metaphor works for most individuals as a straightforward and very well alighned with the definition."	
Definitional alignment	"It is a nice picture to muscle tension, but other physiological and psychological signals might not be entirely covered"	Not capturing the (full) psychological aspect/ not capturing the (full) physiological aspect
Definitional alignment	"the aspect of anxiety is not very well captured in this picture"	Not capturing the (full) psychological aspect
Definitional alignment	"The scale makes sense to me, while it is a relatively new way of measuring tension. I'm not so sure about the scale validity."	Other
Definitional alignment	"heart rate and sweating is not really convincingly presented, but the concept "tension" as such is recognizable"	Not capturing the (full) physiological aspect
Definitional alignment	"This single-item scale do not incorporate the different levels of tension (i.e., psychological and pysiological effects). Nevertheless, I think the scale is valid in measuring tension in a rather intuitive way."	Too simplistic to capture the whole construct
Definitional alignment	It fits the word tension, but I don't understand how it fits the definition of psychological tension. I also don't understand why this measure is better than the established SAM pictorial scale	Not capturing the (full) psychological aspect/ no additive value beyond Self-Assessment Manakin
Relevance	"The content is highly relevant for measuring the construct of tension, especially its psychological aspects. The rubber band metaphor provides a clear, relatable, and simple way to gauge subjective tension. While it may not fully capture all physiological aspects of tension, the content	Mentioning practicality/ addition to design

	is useful for most practical purposes and would work well in various assessment settings. If the scale's primary focus is psychological tension, then its relevance is excellent; for physiological tension, additional representations might enhance its comprehensiveness.”	
Relevance	“Nice, very easily relatable measure. Some small edits could be made maybe to distinguish the high points of the scale better (the "red spot" in the fingers is maybe more visible in 4 than in 5 point, although I see why this is so, another point is that the wrinkly lines above the rubber band could be better distinguished - just looking at the lines without considering the other elements in the pictures I don't "feel" a big difference in the tension from 3 to 5. Maybe this is due to the tension in the band actually most going from left to right, whereas the wrinkly lines imply that the band would vibrate up and down?)”	Changing the design
Relevance	“I think it's fairly relevant, although I'm not sure if it would add much beyond the "arousal" dimension of the Self-Assessment Manakin: https://doi.org/10.1016/0005-7916(94)90063-9”	No additive value beyond Self-Assessment Manakin
Relevance	The picture is more like a physical tightness than an emotional tension. We may need further words to associate picture with introspective reflection.	Not capturing (full) psychological aspect/ expanding the instructions
Comprehensiveness	“There is nearly no shorter way to measure tension.”	Mentioning practicality
Comprehensiveness	“The physiological element (heart rate, sweating) are not fit with this image to me. Maybe if there would be pictures with	Not capturing the (full) physiological aspect/ changing the design

	straining and releasing the rubber it would be easier to imagine the physiological part of the strain"	
Comprehensiveness	"see above - the picture is more on tension as a state, but not really on arousal as a dynamic process"	Too simplistic to capture the whole construct
Comprehensiveness	"The purpose of having a pictorial scale seems to be simplistic. It has some merits, while is also lacking in the comprehensiveness."	Too simplistic to capture the whole construct
Comprehensiveness	"The scale is moderately comprehensive in covering the construct of tense arousal. It does an excellent job of capturing the psychological aspect of tension and provides a clear, intuitive way to measure varying levels of subjective tension. However, it is less comprehensive in addressing the physiological aspects of tense arousal, which are an essential part of the construct as described in the definition. To be fully comprehensive, the scale could benefit from including more explicit indicators of physiological responses to tension and perhaps better representing the extreme ends of the arousal spectrum. Despite these limitations, the scale remains useful and practical for quickly assessing tension, particularly in situations where subjective psychological tension is the primary focus."	Too simplistic to capture the whole construct
Comprehensiveness	"Maybe some common-sense components of psychological tension might be overshadowed by an image mostly reflecting the physical dimension of tension."	Not capturing the (full) psychological aspect
Comprehensiveness	"Good comprehension. I think it covers the psychological side really well. For the	Not capturing the (full) physiological aspect

	physiological side, the tension in muscles may be most captured - I'm not sure if heart rate etc. relates as strongly to the rubber band analogy"	
Comprehensiveness	"It feels like it captures more tense/on edge aspects to to me than a more chaotic emotional anxiety/nervousness (like the SAM). Maybe the wavvy lines can be more chaotic like in different directions?"	Not capturing the (full) psychological aspect/ changing the design
Comprehensiveness	"Sweat could maybe be displayed somehow, too."	Addition to design
Instructions	"I think the instructions are totally clear and no further improvments are needed."	Acknowledging clarity of the instructions
Instructions	"No, very clear"	Acknowledging clarity of the instructions
Instructions	"I don't have other feedback, the instructions are clear to me"	Acknowledging clarity of the instructions
Instructions	"maybe you could use also other images not realted to the stretch that as two end seems also related to something else or someone. so an image of just one object that change"	Addition to design
Instructions	"From a psychometric point of view, a seven-step measurement could be advantageous over a five-step measurement, especially if it is a single-item scale. Perhaps a gradation could be made between 1 and 2 and between 4 and 5?"	Expanding the scale by items or options
Instructions	"This is a very interesting way to visualize tension! I have a thought (not sure it would be correct, but decided to share for your consideration). I interpret that the pictures with the "waves" represent tension, and those without waves lack of tension. I wonder if it could be useful to have one more picture as "2" where the band is slightly hanging. Less hanging than 1 but looser than 3. Then, you could have midpoint (3) as	Changing the design

	the straight band with no waves, as a neutral state. And then two states with waves as your 4 and 5. But again, this is just a thought. Your scale is probably good enough as it is. Good luck with your work!"	
Instructions	"perhaps you can ask a few questions to comprehensively measure tension."	Expanding the scale by items or options
Instructions	"No. Just congratulate the researchers for the idea, it seems a nice way to overcome some limitations of regular questionnaires"	Mentioning practicality
Instructions	"I think the term worried is useful to use as well as anxiety, as anxiety can imply a more extreme affective state"	Expanding the instructions
Instructions	"Overall I find the current pictorial scale a good proxy measure. I feel that addition broadening the representation (e.g., facial expression matching the degree of tension) could enhance construct validity, but they may hinder the measure feasibility and agility as well, so I would keep it as it is."	Addition to design
Instructions	"An example for stressor can be added if it is possible, maybe one word can added to the picture for orientation, like a scale (less, strong etc.)"	Expanding the instructions
Instructions	"Instructions are clear to me. Just wondering if this picture is revelant to any culture or country"	Other/ acknowledging clarity of the instructions
Instructions	"Maybe it should be mentioned in the instructions that an unpleasant state of tension is assessed and therefore described by the rubber band illustration. This is not intuitively clear as tension can be perceived positively as well (being awake, energetic, feeling strong or resilient)."	Expanding the instructions/ improving clarity of the instructions

Instructions	“I think deciding between options 2-4 might be difficult.”	Changing the design
Instructions	“Maybe it would be helpful to display a face in addition to the hands. Additionally, I feel like the not-tense-at-all picture looks a bit weird. Maybe there is a better way to display a relaxed rubber band?”	Addition to design/ changing the design
Instructions	“What suddenly comes to my mind is that you may add a phrase in the beginning like “ which of the following drawings represent your inner feeling” or something like thisl”	Expanding the instructions
Additional feedback	“No, it's enough clear”	Acknowledging clarity of the instructions
Additional feedback	“I'm just wondering whether the condition of a torn rubber band should also be shown, or whether this is already a different construct.”	Addition to design
Additional feedback	“This is a great idea (especially for applied research)! Even though I mentioned the SAM before as a comparison, this variation could be an improvement on it by integrating some of those prior elements. Maybe even having this held in front a person's chest to show it's an internal state?”	Addition to design
Additional feedback	“In the instructions a ”stretched” rubber band is mentioned . However, in the first category of the scale the rubber band is not stretched at all. By describing a stretched rubber band, participants can be primed concerning their own state of tension.”	Other/ improving clarity of the instructions
Additional feedback	" Were facial expression considered to highlight the psychological effects? The stretching of the rubber band could also be pleasurable”	Addition to design

Appendix B. Document on the use of AI in this thesis

In this thesis, I used two AI models, namely Microsoft Copilot (2025) and OpenAI (2025). I used these models to fine-tune my own work (language correction and language assistant) and as general sparring partners for brainstorming. When in doubt about word choice or grammar, I used these models to check my sentences and apply their tips to improve them. Also, when I wanted to know general information about a certain topic, which I could have found myself if needed, I asked the AI models. I didn't use AI in a way that interfered with my own learning process, and merely used it as a tool to do tasks which I also could have done otherwise if I had put time into it, like grammar questions or general knowledge.

References

Microsoft. (2025). *Microsoft Copilot* [AI-powered productivity tool]. Microsoft. https://www.microsoft.com/en-us/microsoft-copilot

OpenAI. (2025). *ChatGPT (GPT-4)* [Large language model]. OpenAI. https://openai.com/chatgpt