Alternative Methods to Evaluate Divergent Thinking Tasks – Effect of Dimension and Procedure on Face Validity and Workload

Julynn Kittel

Master Thesis – Work, Organizational and Personnel Psychology

S4684133
February 2022
Faculty of Behavioral and Social Sciences
University of Groningen
Examiner/Daily supervisor:
Dr. Eric Rietzschel

Abstract

The widely accepted bipartite definition of creativity, including originality and usefulness is often operationalized as such to evaluate divergent thinking (DT) tasks. Based on prior research, showing that these two subdimensions interact with each other to constitute creativity, the (face) validity of their operational separation for measurement was questioned. Another issue to be tackled was the workload that stems from scoring ideas (resulting from DT tasks) individually and twice. To examine methods decreasing workload, the common procedure was compared to the snapshot scoring technique. Furthermore, to investigate possible effects of another operational dimension definition on face validity, a combined 'original usefulness' dimension was proposed and compared to the separate dimension approach. Behavioral science university students (N = 198) took part in an online experiment (survey) with a 2x2 factorial design. They either rated ideas or idea sets (snapshot scoring) and either according to the separate dimensions or the combined 'original usefulness' dimension. The combined dimension approach showed greater outcome-related face validity reinforcing prior findings on the interaction effect of the two subdimensions and opposing their independent operationalization. Individual idea scoring was perceived as more accurate as compared to the snapshot scoring and interrater-reliabilities strengthened this stance. However, measuring outcome-related face validity reinforced the promising nature of the snapshot scoring technique. Taken together, future research should further explore the potential of the snapshot scoring technique and ways to increase inter-rater reliability when utilizing a combined dimension approach.

Keywords: creativity measurement, divergent thinking tasks, snapshot scoring technique, operational definition, face validity, workload, interrater reliability

Alternative Methods to Evaluate Divergent Thinking Tasks – Effect of Dimension and Procedure on Face Validity and Workload

Creativity has helped human kind in every era to move forward, to develop, and to survive. Creativity has a long history if you think of Neanderthals which have developed their artifacts with the help of creativity or nowadays the start-up or research teams who develop innovative solutions for sustainable energy supply with the help of creativity. As a research subject, creativity has gained more and more attention in recent decades. Accordingly, many theories and different tools to measure creativity have been established over time. Wide consensus has been reached across different measurement methods of creativity. However, given the common pragmatic approach, meaning researchers in the domain have focused mostly on developing and understanding creativity; a lack of testing the validity on the ideas about creativity can be seen (Sternberg & Lubart, 1999). That is, well-known techniques (of which some will be outlined in the present study) have been adopted for many decades but questioning the status quo and scrutinizing the standard path has fallen by the wayside. Consequently, with the aim to improve creativity measurement, this paper will tackle issues of common measurement methods by comparing them to new alternatives and hence questioning their validity and utility. Specifically, in this study, I will explore two alternative measurement methods to evaluate creative ideas and compare them to a more common, "standard" operational approach. The measurement methods will be compared by looking at their workload, face validity, and interrater reliabilities.

Creativity and Creative Ideas

The study at hand investigates measurement methods for creative ideas. Ideas are predominantly measured by raters' judgements. But what do they base their scorings on?

According to the most widely agreed conceptual definition, creativity incorporates two dimensions; namely *novelty* and *usefulness* (Barron, 1955; Runco & Jaeger, 2012; Stein,1953). As defined by Feist (1998) for something to be creative it should be novel-original and useful-adaptive. Something is considered *original* if it has not been seen or mentioned within a given context before (Stein, 1953). Focusing on ideas, a creative idea is original in that it is seldomly found within a set of ideas (Litchfield et al., 2015). The second component *usefulness* completes the conceptual definition because something should be adaptive to reality and feasible to be called creative (Barron, 1955; Ochse, 1990). As an example, the idea to present lectures in a more original manner by hanging a lot of tinsel in the lecture hall would be something completely new to everything seen before but it would most likely not add any value to the event. Thus, the idea would be original but not useful and hence not creative. Thus, according to the widely agreed definition, ideas are only considered creative if they are both original and useful (Mackinnon, 1965). This bipartite conceptualization including originality and usefulness has been applied and recognized for many decades and is also referred to as the ‘standard definition’ of creativity (Runco & Jaeger, 2012).

Besides this wide consensus of the conceptual definition, there are multiple ways to study creativity. A basic distinction that can be made is between the so-called four P’s of creativity – Person, Process, Press, and Product (Rhodes, 1961). The *person* perspective deals with individual characteristics of a person that are associated with creativity. For example, personality traits of eminent artists have been studied to determine possible trait indicators for creativity (Feist, 1998). The *processes* perspective, on the other hand, investigates the actual experience of the creative process (Kaufman et al., 2008). Structural neuroimaging can be used, for instance, to find out which specific parts of the brain are activated while someone solves a task requiring creative thought (e.g., Cousijn et al., 2014). Furthermore, the *press*'s

point of view focuses on the environment that stimulates creativity (Kaufman et al., 2008). Researchers, for instance, classified aspects of the work environment, like freedom, recognition, adequate resources, and challenging work as conducive for creative output (Amabile & Gryskiewicz, 1989). Lastly, the *product* approach – focus of this study – revolves around the nature of creative output (Kaufman et al., 2008). This comprises all kinds of creative outputs or responses, like for instance a poem, a story, or as in the study at hand: the property of ideas.

Overall, all four P's often heavily rely on some sort of product (i.e., creative output) measure but research on the measurement of creative output itself is rather scarce. A few studies, however, investigated the measurement of creative ideas. One study by Runco and Charles (1993), for instance, operationally applied the above mentioned 'standard definition' of creativity and judges were asked to sort idea pools based on their originality, appropriateness (similar to usefulness), and creativity. In their study it was shown that idea pools' high originality scores were strongly correlated with high creativity scores but appropriateness scores were less predictive of creativity. Similarly, a study of Diedrich et al. (2015) testing the operationalization of the 'standard definition' for idea evaluation, showed that the usefulness component was only predictive of creativity in highly original ideas but not in ideas scoring low on originality (Diedrich et al., 2015). Another example of a study on idea evaluation was done by Mouchiroud and Lubart (2001) in which different scoring methods to evaluate originality were adopted and compared with each other. It was found that the different scoring methods hugely affected the outcomes and hence the interpretations of the results. Thus, it was shown that different operational definitions lead to different creativity indexes and scholars should therefore consider these variations.

Measurement of Creativity and Creative Ideas

Most of the time, creativity is measured using rater-based judgments. This operational method is based on the notion that evaluation of creativity is inherently subjective (Amabile, 1996, 1982). According to Amabile (1982), something is considered creative to the extent that people agree that it is. Thus, consensus among raters is usually calculated as a baseline reliability measure for creativity assessment.

Based on the acknowledgement of subjectivity in creativity and pointing out the importance of the judges, Amabile (1982) introduced the consensual assessment technique (CAT). This method aims at assessing domain-specific products such as poems and short stories (Kaufman et al., 2008). Essential to this method is that *experts* in the field under consideration evaluate the created products. According to Amabile (1982), creative products should be evaluated in terms of consensus between experts because the concept of what is creative is largely shared among experts as tacit knowledge. As an example of the application of the CAT, participants of a study could be asked to write a poem which then domain experts (i.e., proficient poets) would evaluate. Importantly, every expert rater would evaluate each of the created poems to be able to gather consensus among the judges to constitute the overall evaluation of the poems.

A subjective rating approach is also mostly used for the evaluation divergent thinking tasks. The divergent thinking task is one of the most widely applied creativity assessment tools and it requires, like the name says, divergent thought (Kaufman et al., 2008). Divergent thinking, coined by Guilford (1967) as a core skill of creativity, refers to the ability to come up with a variety of different ideas when faced with a problem. There are different kinds of tasks to assess divergent thinking, but they all have in common that people generally have to come up with different ideas based on some kind of stimulus like a statement, an object or problem to be solved. For example, participants of a study could be presented with a question like ‘How can we increase attractivity of the city center?’ and based on this question,

respondents are asked to come up with as many creative ideas as possible to provide solutions to the posed question.

Subsequently, subjective judgements of raters are used to evaluate the generated ideas of the divergent thinking task. Most of the times they are evaluated in accordance with the 'standard definition' of creativity, thus, according to their originality and usefulness level or according to a combination of both (Forthmann et al., 2017; Reiter-Palmon et al., 2009; Runco & Charles, 1993). More specifically, the translation of the conceptual 'standard definition' to the operational definition means that raters are usually instructed to assign two separate scores for each idea, one for the novelty and one for the usefulness dimension (Diedrich et al., 2015). Every idea's creativity level is then defined by, for example, the sum of these two scores and the average across judges (see e.g., Rietzschel et al., 2007).

Two Issues

Despite the widespread usage of the above mentioned and comparable procedures to assess creative ideas, there are two issues at hand that will be tackled in the current study. The first issue concerns the procedure's labor intensity, whilst the second revolves around its' face validity.

Labor-intensity

It is exhausting and costly in nature to rate every idea of every participant individually (Shaw, 2021). All the more, applying the 'standard definition' to evaluate the creativity levels of ideas, thus, scoring each idea twice, once for originality and once for usefulness, results in huge amounts of scorings. Around 50,000 ratings can build up in a study with a large sample, several tasks, and multiple raters (Silvia et al., 2009). One can imagine how many hours, coffee, and how much concentration it takes to get these ratings done. Accordingly, it has been found that devoting these enormous amounts of hours on the rating task can result in rater fatigue (Cseh & Jeffries, 2019). According to Cseh and Jeffries (2019), in a study of

Amabile (1982, Study 1), for example, a link between time spent on the scoring task and interrater agreement was suggested. Thus, the time and energy demanding nature of these measurement methods might also lead to scoring inaccuracies. Finally, this investment requires extra motivation if the divergent thinking task is only one of the research measures and, as in some studies, creativity is only a secondary variable (Shaw, 2021).

To reduce the workload of divergent thinking task assessments, Silvia et al. (2009) have proposed alternative methods like the *snapshot scoring* (i.e., ideational pool scoring, where participants assign a score to a set of ideas rather than to each idea separately). Snapshot scoring is thought to alleviate workload by evaluating idea sets of participants as a whole instead of assessing each idea in a set individually. In the past, a few scholars used ideational pool scoring to evaluate divergent thinking tasks (Mouchiroud & Lubart, 2001; Runco & Mraz, 1992). That is, to give one overall score to the whole set of ideas of a participant instead of considering each idea of every participant one by one. In an earlier study of (Runco & Mraz, 1992), adolescents participated in divergent thinking tasks, and college students scored the creativity level of each participant's set of ideas. The holistic scores proved good reliability but indications of validity properties were not obtained. Later on, a comparable study was done by Mouchiroud and Lubart (2001) in which raters were asked to assign overall scores to idea sets generated by children. One group of raters scored the idea sets based on 'creativity' and the other group of raters scored them based on 'originality'. Again, the scores demonstrated high reliability, but no validity values were gathered.

Based on the promising findings of these prior studies, Silvia et al. (2009) picked it up for further psychometric investigation and termed it *snapshot scoring*. Beside the promising reliability findings, they sought to examine the validity of the ideational pool scoring technique. Thus, in their study, the snapshot scoring (i.e., ideational pool scoring) method was

adopted and its' concurrent validity was examined. Concurrent validity is the examination of whether the results of one test are related to the results of another criterion-related test taken at the same point in time (Carmines & Zeller, 1979). In their study it was therefore determined whether the snapshot scores (i.e., overall idea set score) of persons idea sets correlated with results on specific personality tests. As an example, it was examined if results of the snapshot scoring correlated with scores on openness to experience (a personality trait in the Big Five personality test) as it has been repeatedly proven that this personality trait is usually higher in creative people (Feist, 1998). Results of their analysis showed that openness to experience did indeed account for a substantial amount of variance in the snapshot scores, indicating strong concurrent validity (Silvia et al., 2009).

According to this sum of findings, reducing workload by scoring ideational pools rather than individual ideas has been shown to be a promising approach, as psychometric properties have been shown to remain high. On the contrary, however, it has been found that assigning a single score to an individual idea rather than assigning a single score to a collection of distinct ideas can result in greater mental workload (Forthmann et al., 2016). More precisely, the effect of complexity on mental workload was found to be stronger when scoring idea sets than when scoring single ideas because idea sets generally display more variance in complexity. As a result of this finding, it was indicated that scoring ideational pools significantly reduces the number of ratings but may also significantly increase the mental workload per judgment which is likely to be a source of interrater-disagreements (Forthmann et al., 2016). Thus, given these contradictory findings on the possibilities of the snapshot scoring technique, it is important to conduct more research on it.

Operationalization of Subdimensions

The second issue with the aforementioned common procedure for rating divergent thinking tasks is how the two subdimensions of creativity are operationalized. There are

debates and research findings that argue against the independent operationalization of the two subdimensions. Overall, these prior findings indicate that this kind of measurement method could result in low levels of face validity. That is, the subjective perception of the validity of a method (Mosier, 1947).

Apart from widespread acceptance of the 'standard definition' of creativity, it is debated whether a single definition can be applied to all instances or forms of creativity. As a result, researchers are continuously challenging and expanding the predominant conceptualization. For example, Madjar et al. (2011) illustrated that ideas with low novelty are not necessarily uncreative ideas. They differentiated between two kinds of manifestations of novelty in creative ideas, namely, “suggesting radically new ways” (radical) and “adapting existing ideas” (incremental). This differentiation showed that, depending on the needs and contexts, ideas with low novelty can be as (or even more) beneficial as ideas with high novelty. Following this, Litchfield et al. (2015) distinguished between two subdimensions of the usefulness component, namely the *value* and *feasibility* of ideas. According to their thesis, a 'low-hanging fruit', for instance, is an idea that is low in novelty but high in value and feasibility. On the other hand, a highly novel idea is referred to as a radical idea if it is high in value but low in feasibility (Litchfield et al., 2015). Another study examined the effect of originality and usefulness on a product's word-of-mouth. Among others, it was discovered that highly original products with low feasibility induced more negative word-of-mouth than less original products with the same level of feasibility (Moldovan et al., 2011). Overall, these studies demonstrate that focusing exclusively on high levels of originality and usefulness, and separating these two dimensions, does not tell the entire story of the creativity construct.

There is further evidence opposing the common independent and additive operationalization of the two subdimensions and suggesting low face validity of this method. According to anecdotal evidence (E. F. Rietzschel, personal communication, April 22, 2021),

scoring originality and usefulness (here: feasibility) separately sometimes leads to ideas being classified as highly creative which, upon closer consideration, do not seem to be particularly creative. For instance, to keep students more awake during a class, they could be asked to stand up every 15 minutes. This might be something very new and very feasible but overall, not necessarily creative. Additionally, the relationship between novelty and usefulness has been demonstrated to be negative in creative ideas (see e.g., Runco & Charles, 1993). Furthermore, a study by Diedrich et al. (2015) found that usefulness and novelty do not contribute to creativity in the same way and that they interact with each other. That is, the degree to which usefulness contributes to creativity depends on the magnitude of novelty (Diedrich et al., 2015). Thus, usefulness seems to be particularly important in highly novel ideas but less so in ideas with low or medium novelty. Taken together, it has been shown that the two dimensions interact in quite complex ways which leads to the assumption that separate operationalization of them may not do justice to the measured creativity construct. Thus, operationalizing the subdimensions in an independent manner can result in ideas being labelled as ‘highly creative’ which after closer examination might not represent a highly creative idea. Thus, looking at the face validity, I expect raters to subjectively evaluate the measurement method with low validity when they are scoring the two subdimensions separately.

Besides the typical separation of the two underlying dimensions, researchers sometimes instruct raters to score on the overall construct 'creativity' (e.g., Benedek et al., 2013; Silvia et al., 2008). To emphasize the recognized subdimensions of creativity, however, I propose an alternative operational definition which explicitly incorporates both dimensions into one single construct. In a study of Withagen and van der Kamp (2018), they outlined the assumption that patterns which are to be explained, already exist beforehand, although in an abstract from. Accordingly, to take away the abstractness, I propose the alternative to evaluate

ideas according to '*original usefulness'*. That is, raters should ask themselves if the ideas are useful in a novel way. This alternate dimensions explicitly takes into account the interaction effect of the subdimensions as raters are required to assign one single score rather than two separate ones. This approach can also be compared to a conceptual definition of Bruner (1972) who referred to creativity as being an *effective surprise*. Thus, by using this term, he also explicitly emphasized the two subdimensions considered to constitute creativity. That is, something surprises (originality) and is effective at the same time (usefulness).

By explicitly incorporating the two subdimensions into one, I expect the subjective perception of the measurement method's validity (face validity) to be higher as compared to the separate operationalization. Apart from pragmatic and statistical validity, the subjective assessment of validity is equally worthwhile (Mosier, 1947). Unfortunately, this form of validity receives scant attention in research as it is often dismissed as trivial (Andres, 2012). Even though, it can be beneficial to test if a measurement method not only *is* valid but also *appears* valid (Mosier, 1947). One of the few studies examining face validity in creativity research was done by Harris (1960), in which creativity of engineers was measured and face validity was examined by asking testees: 'Do you think that the test you have just taken can measure creativity in engineers?'. Thus, they were directly asked about their perception of the method's ability to measure the construct it was designed to measure. As a result, the testees' responses were used as indicators of the measurement method's face validity. Apart from creativity research, there are other studies, for instance, in the health care setting testing the face validity of questionnaires. In the study of McElroy and Esterhuizen (2017), for example, lay people were asked to review a questionnaire on compassionate communication (with patients) by evaluating its' clarity, simplicity, and suitability. Seeking the subjective perception was in this case ought to serve as a pre-contemplation to examine the clarity and fit of purpose of the questionnaire (McElroy & Esterhuizen, 2017). Thus, face validity can serve

as an additional validity check by examining not only the statistical validity but also the subjective perception of the validity.

The present study

Tackling the mentioned issues of common ways of divergent thinking task evaluation, the study at hand will test the effectiveness of 1) the snapshot scoring method as opposed to individual idea scoring and 2) the combined 'original usefulness' dimension as opposed to separate dimension scoring. The effectiveness of these alternative methods for idea evaluation will be assessed in terms of:

Workload. The workload of the measurement method will most likely benefit (i.e., be lowered) from the alternative methods of scoring idea sets and applying a combined dimension as operational definition. Thus, workload is supposed to be significantly reduced, especially applying the snapshot scoring technique.

Face Validity. Given the research counteracting the operational separation of the subdimensions originality and usefulness, I expect that especially the adoption of the combined dimension 'original usefulness' for idea evaluation will strengthen the face validity (the subjective validity perception) of the measurement method.

Interrater Reliability. Lastly, Interrater agreements as the core of creativity measurement serves one of the most important psychometric values to check for the reliability of an assessment method. With the aim to improve measurement methods, it is crucial that interrater reliabilities remain high.

Methods

Participants

Using the SONA-System of the University of Groningen and snowball sampling, a total of 197 participants (1 non-binary, 151 females, 44 males) took part in the study. One participant's data set was deleted because not all fields were filled out. Students took part

through the SONA-System and received credits as part of a mandatory course at the university. The content of the study and the informed consent were reviewed and approved by the ethical committee of the University of Groningen.

Participants were invited to provide their age in terms of age brackets ranging from ’17 or younger’ to ’60 or older’. Most (76 %) participants were between 18-20 years old and second most participants (21.4 %) were 21-29. Based on an a priori power analysis using ANOVA, 190 participants were required to be able to observe a medium to large effect size of 0.30 with a power of .80.

Design and Procedure

The study was implemented as an online experiment (i.e., a survey) with a factorial 2x2 design. Participants were redirected to the Qualtrics platform for the survey to begin. Participation was voluntary. Prior to the survey beginning, a consent form was provided and participants could decide whether they wanted to take part in the study or not. The survey was anonymous and only age brackets, educational background, and gender were asked for. After agreeing to the consent form, the actual survey started and took participants about 10 minutes to complete. At the end of the survey, participants were provided a debriefing letter explaining the intention and theoretical background of the study.

Participants were randomly assigned to one of the four different conditions. In each of the conditions, they had to assign scores to various ideas or idea sets based on their creativity level. Ideas or idea sets were always presented in a random order, so that every rater was presented with a different order. The ideas, participants were provided with for the rating, were gathered in a former study about maintaining or improving health (Rietzschel et al., 2007). After rating the ideas, they were presented with a part of their rating outcome and were asked to indicate their agreement with the outcome to represent what it was supposed to measure. Further, they were asked a few questions to indicate the clarity and the perceived

workload of the measurement technique. Additionally, participants were asked to indicate to what extent they perceived that the measurement technique they were using covered the construct it purported to cover.

Materials

Ideas. 28 ideas were chosen to be rated in this study. The ideas were chosen based on their previous study's scores (see Rietzschel et al., 2007). Scorings of the previous study were based on two distinct dimensions, namely originality and feasibility. To ensure sufficient variance between the ideas, they were selected based on their distinct values of originality and feasibility. Scores of the previous study were categorized in high, medium, and low levels. Ideas were accordingly chosen in a manner to ensure equal representation of each category (see Appendix A)

Idea Sets. Seven idea sets were created from the 28 selected ideas. The idea sets were also created in a manner to ensure that different combinations of the categories were represented (see Appendix B). Thus, to ensure variance across the idea sets, they were, set up to compose particularly high, medium, and low valued ideas based on their creativity scores (which were based on the calculation of adding their originality and feasibility scores).

Independent Variables

Procedure. Dependent on the assigned condition (individual ideas or snapshot scoring), participants either scored ideas individually or they scored sets of ideas. In other words, in the 'individual ideas' condition, participants were asked to rate every single idea they were presented with separately. In the 'snapshot scoring' condition, following the procedure of Silvia et al. (2009), participants were asked to rate sets of ideas, providing one score for each set of ideas. Participants had to rate a total of 28 ideas. According to the conditions, they were either asked to rate them individually or these 28 ideas were compiled into seven sets; with four ideas per set. In that case, they were rating seven sets of ideas

instead of 28 ideas individually. The ideas they were presented with were drawn from a former study of Rietzschel et al. (2007). The study included a divergent thinking task in which participants had to come up with as many creative ideas as possible on how to maintain or improve health.

Dimension. To compare different dimensional approaches to creativity measurement, two different conditions were created. Participants either rated the ideas (or: idea sets) based on one dimension, namely the merged 'original usefulness' term or in the other condition based on two dimensions; first on their originality and afterwards on their usefulness level. In each condition, they were instructed more thoroughly about the operational definition and measurement technique they were asked to apply.

Dependent Variables

To evaluate the ideas, participants were provided a 5-point Likert Scale. All other measures to evaluate the dependent variables were based on a 5-point Likert Scale as well.

Workload. To measure the perceived workload of the measurement method, a 3-item scale was used. Firstly, the 1-item Rating Scale Mental Effort was used (Zijlstra & Doorn, 1985): 'Using this rating procedure required a lot of work'. To simplify the scale and to make it comparable for data analysis, the scale was changed from a 0-150 to a 5-point Likert scale ranging from 1 ('Not at all') to 5 ('Extremely'). Secondly, participants were presented with two additional items: 'Using this rating procedure was difficult' and 'Using this rating procedure was easy to understand'. Further clarification was provided about the meaning of 'difficult' and 'a lot of work' within the context of the study. Internal consistency for the total 3-item scale was .62.

Face Validity. Face validity is a complex construct including subdimensions such as accuracy, acceptance, relevance, perspective, and accurate completion rate (Thomas et al., 1992). Thus, to measure different angles of face validity, two different variables were applied.

The first variable was a scoring outcome-related face validity measure including three items. Participants were presented with part of their rating outcome, displaying the top three ideas (or idea sets) according to their rating (that is, the ideas or idea sets that they gave the highest ratings to). Presented with a part of their rating outcome, they were requested to evaluate the extent to which they agreed with the outcome to represent what was supposed to being measured (either creative ideas or creative idea sets). To get a thorough picture on their (dis-)agreement with their rating outcome, they were asked three similar but distinct questions. The three items of the scale were phrased as follows: 'To what extent do you feel that these ideas are actually creative?', 'To what extent do you feel that these ideas deserve to be in the top 3?', and 'To what extent do you feel that these ideas were the most creative ones?'. Participants gave their responses on a 5-point scale from 1 ('Not at all') to 5 ('Extremely'). Internal consistency for the 3-item scale was .68. The second face validity variable was a more direct measure to collect explicit opinions on the perceived suitability of the measurement method they have been using. Accordingly, the 1-item face validity scale of Nevo (1985) was used: 'To what extent do you perceive this measurement technique as suitable for the given purpose?'. Participants gave their responses to this item on a 5-point scale (1 = This measurement technique is irrelevant and therefore unsuitable, 2 = This measurement technique is inadequate, 3 = This measurement technique is adequate, 4 = This measurement technique is very suitable for that purpose, 5 = This measurement technique is extremely suitable for the given purpose).

Interrater Reliability. According to the definition that something is creative to the extent that people agree that it is (Amabile, 1982), consensus across raters builds the basis for subjective creativity judgements. The Intra-class correlation coefficient (ICC) was used to assess the interrater reliability. The ICC indicates how well a measure can distinguish between scores indicated by two or more raters (Kottner et al., 2011; Liao et al., 2010). Due to

the fact that raters were consistent for all items in each condition and a random effect of ideas but a fixed effect of raters was assumed, the ICC (3) for two-way mixed models with consistency was calculated (Landers, 2015). Furthermore, to get information about the accuracy of a single person, the single measure was examined.

Results

Preliminary Analysis

Descriptive statistics and bivariate correlations for three out of four dependent variables (Nevo face validity scale, outcome-related face validity scale, and workload) were calculated and are illustrated in Table 1. Due to the group-level nature of the fourth dependent variable (interrater reliability), it could not be included in the correlations. As expected, the analysis showed that the two dependent face validity variables correlated significantly with each other (r = .238, p < .001). Workload did not correlate with either of the face validity measures.

Assumption Checks

Multivariate ANOVA. Based on the positive correlation of the two dependent face validity scales, a multivariate ANOVA was chosen to examine the effect of dimension and procedure on face validity. Before conducting the MANOVA, the assumption of equivalence of covariances was tested. Results were not significant (F (0.76), p = 0.65) which means the assumption has not been met and adopting a MANOVA was affirmed. Checking for normality, the Shapiro-Wilk was significant (p = <.001) for both face validity measures, showing that both variables were not normally distributed. Accordingly, results of the MANOVA should be interpreted with caution. However, using a linear regression analysis, calculation of the Mahalanobis distance showed multivariate normality for the face validity measures as the critical value of 13.82 was not exceeded (Mahalanobis distance = 8.57).

Furthermore, homogeneity of variances was tested for both face validity measures and were satisfied by means of the Levene's F test.

ANOVA. Before conducting the ANOVA for the second dependent variable (workload), the assumption of homogeneity of variances was tested and satisfied by means of the Levene's F test.

Statistical Tests

Workload. To see if workload differed between conditions, a two-way ANOVA was conducted. In contrast to expectations, results of the analysis showed that neither procedure (F (1, 192) = 0.13, p = 0.72), nor dimension (F (1, 192) = 0.95, p = 0.33) had a significant effect on workload. There was also no significant interaction effect of the two independent variables (F (1, 192) = 0.03, p = 0.86) on workload. Given the large differences in the number of ratings between conditions, it can be assumed that the workload scale in use did not capture what it was supposed to capture.

Face Validity. To explore the effect of both independent factors, namely dimension and procedure, on both dependent face validity variables, I conducted a multivariate ANOVA. The results showed that both 'dimension' (Wilks-Lambda = .031) and 'procedure' (Wilks-Lambda = .003) had a significant effect on the face validity measures. However, the MANOVA results showed that there was no interaction effect of the two factors (Wilks-Lambda = .161). Looking at the independent effects of the two factors on the face validity measures, it showed that the effect of 'dimension' was significant for the outcome-related face validity scale (F (1, 192) = 6.02, p = .023). Accordingly, participants reported higher outcome-related face validity when scoring with the combined dimensions (M = 3.39, SD = .85) than when scoring with the separate dimensions (M = 3.09, SD = .87), suggesting that scoring ideas according to the combined construct results in outcomes that better reflect the creativity construct than scoring ideas according to the separate dimensions.

There was no significant effect of 'dimension' on the Nevo face validity scale (F (1, 192) = .26, p = .608).

The effect of 'procedure', however, was significant for the Nevo face validity scale (F (1, 192) = 9.25, p = .003). Participants reported higher face validity of the Nevo scale when scoring ideas individually (M = 3.12, SD = .78) than when scoring idea sets (M = 2.78, SD = .81), suggesting that scoring the creativity level of ideas one by one appears more accurate than scoring idea sets as a whole.

There was no significant effect of 'procedure' on the outcome-related face validity scale (F (1,192) = .68, p = .411).

Interrater Reliability. To examine interrater reliabilities, intraclass correlation coefficients were calculated across all conditions. According to the guidelines (Koo & Li, 2016), values of the intraclass correlation coefficients were in the range of poor interrater reliabilities across all four conditions (all below .5). The highest single measure ICC was .379 for the individual idea scoring with separate dimensions (F (55, 2640) = 30.91, p = .000), suggesting superior reliability for this method. The second highest single measure ICC was .278 for the snapshot scoring with separate dimensions (F (13, 611) = 19.52, p = .000). The single measure ICC was .223 for the individual idea scoring with combined dimensions (F (27, 1269) = 14.79, p = 000). Lastly, the single measure ICC was .211 for the snapshot scoring with combined dimension (F (6, 294) = 14.34, p = <.001). Due to diverging numbers of ratings and distinct measurement foundations across conditions, checking the differences of the intraclass correlation coefficients for significance would exceed the scope of this paper.

Discussion

In the present study, alternative measurement methods (i.e., snapshot scoring technique and combined 'original usefulness' dimension) were explored and compared with a more common approach (individual idea scoring with separate dimension operationalization)

to assess divergent thinking tasks. The face validity of the commonly used method of measuring the two subdimensions, namely originality and usefulness, separately, has been expected to be low. On the other hand, the workload associated with this common technique is fairly high. To address both of these issues, two specific alternative methods were adopted. First, considering the interaction effect of the two subdimensions (Diedrich et al., 2015), they were incorporated into one 'original usefulness' dimension to heighten face validity. Second, the snapshot scoring technique (Silvia et al., 2009) was applied to reduce the workload. Overall, the measurement methods were explored and compared with each other regarding their face validity, workload, and interrater reliability.

Summary and Interpretation

Face Validity. According to the results of the multivariate ANOVA, investigating the effects of dimension and procedure type on face validity, it was shown that scoring according to the combined dimension 'original usefulness' led to stronger face validity (on the outcome-related scale) than scoring the subdimensions separately. Thus, the distinct dimension approaches had significant effects on the outcome face validity measure. On the contrary, procedure had no significant effect on the outcome face validity measure but on the Nevo face validity scale. Accordingly, evaluating ideas individually led to stronger face validity scores (on the Nevo scale) than scoring sets of ideas as a whole (snapshot scoring technique). This means, subjective evaluations of the validity on the outcome of a measurement technique can strongly differ to the subjective evaluation of the validity of a measurement technique itself.

The first scale assessed face validity by using the rating outcome evaluation as an indicator. Raters agreed more with their rating outcome to do justice to the construct when they scored using the combined 'original usefulness' dimension as compared to the separate dimensions. This result is consistent with the expectation that scoring on the combined dimension will better capture the creativity construct. These higher (outcome-related) face

validity results are likely to have emerged because the interaction effect of the two subdimensions was accounted for in the combined dimension approach. Thus, it reinforces the findings of Diedrich et al. (2015), showing that the subdimensions interact with each other and that the effect of one dimension (usefulness) depends on the magnitude of the other dimension (novelty). Overall, it was shown that combining the subdimensions better captured creativity than separating them.

The second scale by Nevo (1985) measured face validity by asking the raters about their perception of the method's suitability for assessing what was supposed to be measured (i.e., creativity of ideas or idea sets). In this vein, results showed that the individual idea scoring method was significantly better ranked than the snapshot scoring technique. Thus, it shows that scoring ideas individually was subjectively perceived as a more suitable measurement technique as opposed to scoring idea sets as a whole. This result could be explained by the findings of a prior study, illustrating that idea sets can have a greater variety of complexity as compared to individual ideas (Forthmann et al., 2016). According to Forthmann et al. (2016), this greater variety of complexity leads to a higher mental workload while judging the ideas. Thus, raters of the present study might have felt incapable to accurately score some of the idea sets due to a high mental workload and might have therefore deemed the snapshot scoring as less accurate.

However, a contradiction was demonstrated; the snapshot scoring (and the individual scoring) with combined dimension had a higher outcome face validity than both the individual and the snapshot scoring with separate dimensions. Thus, when the snapshot scoring was evaluated for its' suitability, it was found to be less valid; however, when the scoring outcome was evaluated, it was found to be superior to the individual idea scoring with separate dimensions. Thus, the latter reinforces the method's previous promising findings (see Silvia et al., 2009).

Workload. Contrary to expectations, neither dimension nor procedure had significant effects on participants' perceived workload. Given the huge differences in the number of ratings that had to be assigned in each condition, it is possible that the applied workload scale did not capture what it was supposed to capture. This result is not necessarily surprising considering that the participants were unfamiliar with the alternative measurement methods and thus had no direct comparison. Alternatively, however, it is possible that the lower workload of having to rate seven sets of ideas instead of 28 ideas was cancelled out by the higher complexity that may have been present in the idea sets (cf. Forthmann et al., 2016). A third possibility is that 28 ideas are not enough for participants to experience a high workload (which is also borne out by the moderate levels of workload reported by participants), and that snapshot scoring therefore did not have a lot of added value.

Interrater Reliability. Across all four conditions, the reliability values were in the lower range. The individual scoring technique with separate dimensions exhibited the highest interrater reliabilities. Likewise, the snapshot scoring technique with separate dimensions led to higher interrater reliabilities as compared to the combined dimension. Thus, it is likely that the separation of the two subdimensions does result in less ambiguity as compared to a combined dimension approach. Similar to findings of Forthmann et al. (2016) on the effect of complexity on mental workload while judging, thinking of two dimensions at once might also lead to more complexity; hence to higher mental workload and more scoring inaccuracies.

Interestingly, however, the low reliability of the combined dimension method contrasts with the method's high (outcome-related) face validity. That means, people subjectively perceived their rating outcomes to do justice to the measured entity, however, they did not seem to agree on what are the most creative ideas or idea sets. This finding strongly emphasizes the subjective nature of creativity perceptions (Amabile, 1982).

Limitation and Future Research

Given the overall low reliability values, there might have been some general ambiguity issues in the rating process. One reason could have been the material, meaning the ideas and ideas sets raters were provided with. The ideas that were chosen from the prior study, strongly differed in terms of formulation, lengths, depths of detail, and level of abstractness. Additionally, keeping context and background information very broadly might have led to different understandings of creative ideas in this broad context. More specifically, the ideas were generated based on the question on how to maintain or improve health, thus the ideas ranged from changing policies to everyday-related actions, such as washing hands more regularly. Therefore, raters might have had different thoughts on which directions to aim for. Thus, having more similar ideas at least in terms of their lengths or their abstractness might lead to higher interrater reliabilities.

Additionally, it would be interesting to investigate possible effects of the explicitly incorporated 'original usefulness' dimension on rating outcomes by comparing it to the use of the sole 'creativity' dimension. Given that some researchers evaluate ideas using the simple 'creativity' dimension (see e.g., Benedek et al., 2013; Silvia et al., 2008), it would be interesting to learn whether explicitly mentioning the two subdimensions in one combined construct shifts the focus of the judgment. Given the widespread agreement on the ‘standard definition’ including these two dimensions (Runco & Jaeger, 2012), comparing these two approaches should not produce significantly different results.

Lastly, the gained insights from examining face validity in the present study, supported the relevance of investigating this type of validity (Mosier, 1947). It was also demonstrated that approaching face validity in different ways widens the insight to a greater extent. That is, using the rating outcome as an indicator for face validity and comparing it to the subjective perception of the method’s suitability provides a clearer picture of the subjectively perceived validity of the method. Thus, results of the present study led to the

conclusion that merely asking about the perceived suitability of a method should always be expanded with a subjective evaluation of the 'agreement' with the rating outcome. Overall, future research on creativity measurement should incorporate those and possibly further face validity measures to be able to draw more comprehensive conclusions. Especially given the simplicity to add those measures to an assessment, it is highly recommended to include these measures in future studies on creativity measurement.

Conclusion

Given the importance of the creative product in creativity research, the lack of research revolving around the creative product itself needs to be filled. Especially considering the issues of common measurement methods, possible ways of improvement should be investigated more thoroughly. Taken together, results of this study support the application of a combined dimension approach for creativity measurement. Furthermore, mixed findings on the psychometric properties of the snapshot scoring technique point to the need to explore its' benefits and disadvantages further. Finally, applying different face validity measures is strongly recommended to expand more commonly used statistical tests of validity.

References

Amabile, T. (1996). *Creativity in context*: *Update to "The Social Psychology of Creativity."* Westview Press.

Amabile, T. M. (1982). Social psychology of creativity: A consensual assessment technique. *Journal of Personality and Social Psychology*, *43*(5), 997-1013. https://doi.org/10.1037//0022-3514.43.5.997

Amabile, T. M., & Gryskiewicz, N. D. (1989). The creative environment scales: Work environment inventory. *Creativity Research Journal*, *2*(4), 231-253. https://doi.org/10.1080/10400418909534321

Andres, L. (2012). *Designing and doing survey research*. Sage.

Barron, F. (1955). The disposition toward originality. *The Journal of Abnormal and Social Psychology*, *51*, 478 – 485. http://dx.doi.org/10.1037/h0048073

Benedek, M., Mühlmann, C., Jauk, E., & Neubauer, A. C. (2013). Assessment of Divergent Thinking by means of the Subjective Top-Scoring Method: Effects of the Number of Top-Ideas and Time-on-Task on Reliability and Validity. *Psychology of Aesthetics, Creativity, and the Arts*, *7*(4), 341-349.

Bruner, J. S. (1972). The conditions of creativity. In Bruner, J. S. (Ed.), *On knowing: Essays for the left hand*. Cambridge, MA: Harvard University Press.

Carmines, E. G., & Zeller, R. A. (1979). *Reliability and validity assessment*. Sage publications.

Cousijn, J., Koolschijn, P. C. d. M. P., Zanolie, K., Kleibeuker, S. W., Crone, E. A., & Lidzba, K. (2014). The Relation between Gray Matter Morphology and Divergent Thinking in Adolescents and Young Adults. *PLoS ONE*, *9*(12), e114619. https://doi.org/10.1371/journal.pone.0114619

Cseh, G. M., & Jeffries, K. K. (2019). A scattered CAT: A critical evaluation of the consensual assessment technique for creativity research. *Psychology of Aesthetics, Creativity, and the Arts*, *13*(2), 159-166. https://doi.org/10.1037/aca0000220

Diedrich, J., Benedek, M., Jauk, E., & Neubauer, A. (2015). Are Creative Ideas Novel and Useful? *Psychology of Aesthetics, Creativity, and the Arts*, *9*(1), 35-40. https://doi.org/10.1037/a0038688

Feist, G. (1998). A Meta-Analysis of Personality in Scientific and Artistic Creativity. *Personality and social psychology review : an official journal of the Society for Personality and Social Psychology, Inc*, *2*(4), 290-309. https://doi.org/10.1207/s15327957pspr0204_5

Forthmann, B., Holling, H., Çelik, P., Storme, M., & Lubart, T. (2017). Typing Speed as a Confounding Variable and the Measurement of Quality in Divergent Thinking. *Creativity Research Journal*, *29*(3), 257-269.

Forthmann, B., Holling, H., Zandi, N., Gerwig, A., Çelik, P., Storme, M., & Lubart, T. (2016). Missing Creativity: The Effect of Cognitive Workload on Rater (Dis-)Agreement in Subjective Divergent-Thinking Scores. *Thinking Skills and Creativity*, *23*. https://doi.org/10.1016/j.tsc.2016.12.005

Guilford, J. P. (1967). *The nature of human intelligence*. McGraw-Hill.

Harris, D. (1960). The development and validation of a test of creativity in engineering. *Journal of Applied Psychology*, *44*(4), 254-257. https://doi.org/10.1037/h0047444

Kaufman, J. C., Plucker, J. A., & Baer, J. (2008). *Essentials of creativity assessment*. John Wiley & Sons Inc.

Koo, T. K., & Li, M. Y. (2016). A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *Journal of Chiropractic Medicine*, *15*(2), 155-163. https://doi.org/10.1016/j.jcm.2016.02.012

Kottner, J., Audige, L., Brorson, S., Donner, A., Gajewski, B. J., Hróbjartsson, A., Roberts, C., Shoukri, M., & Streiner, D. L. (2011). Guidelines for Reporting Reliability and Agreement Studies (GRRAS) were proposed. *Journal of clinical epidemiology*, *64*(1), 96-106. https://doi.org/10.1016/j.jclinepi.2010.03.002

Landers, R.N. (2015). Computing intraclass correlations (ICC) as estimates of interrater reliability in SPSS. *The Winnower 2*:e143518.81744. https://doi.org/10.15200/winn.143518.81744

Liao, S. C., Hunt, E. A., & Chen, W. (2010). Comparison between inter-rater reliability and inter-rater agreement in performance assessment. *Annals of the Academy of Medicine Singapore*, *39*(8), 613-618.

Litchfield, R. C., Gilson, L. L., & Gilson, P. W. (2015). Defining Creative Ideas: Toward a More Nuanced Approach. *Group & Organization Management*, *40*(2), 238-265. https://doi.org/10.1177/1059601115574945

Mackinnon, D. W. (1965). Personality and the realization of creative potential. *American psychologist*, *20*(4), 273-281. https://doi.org/10.1037/h0022403

Madjar, N., Greenberg, E., & Chen, Z. (2011). Factors for radical creativity, incremental creativity, and routine, noncreative performance. *Journal of applied psychology*, *96*(4), 730.

McElroy, C., & Esterhuizen, P. (2017). Compassionate communication in acute healthcare: establishing the face and content validity of a questionnaire. *Journal of Research in Nursing*, *22*(1-2), 72-88. https://doi.org/10.1177/1744987116678903

Moldovan, S., Goldenberg, J., & Chattopadhyay, A. (2011). The different roles of product originality and usefulness in generating word-of-mouth. *International Journal of Research in Marketing*, *28*. https://doi.org/10.1016/j.ijresmar.2010.11.003

Mosier, C. I. (1947). A critical examination of the concepts of face validity. *Educational and psychological measurement*, *7*(2), 191-205.

Mouchiroud, C., & Lubart, T. (2001). Children's Original Thinking: An Empirical Examination of Alternative Measures Derived From Divergent Thinking Tasks. *The Journal of Genetic Psychology*, *162*(4), 382-401. https://doi.org/10.1080/00221320109597491

Nevo, B. (1985). Face Validity Revisited. *Journal of Educational Measurement*, *22*(4), 287-293. https://doi.org/10.1111/j.1745-3984.1985.tb01065.x

Ochse, R. (1990). *Before the gates of excellence: The determinants of creative genius*. Cambridge University Press.

Reiter-Palmon, R., Illies, M. Y., Kobe Cross, L., Buboltz, C., & Nimps, T. (2009). Creativity and Domain Specificity: The Effect of Task Type on Multiple Indexes of Creative Problem-Solving. *Psychology of Aesthetics, Creativity, and the Arts*, *3*(2), 73-80. https://doi.org/10.1037/a0013410

Rhodes, M. (1961). An Analysis of Creativity. *The Phi Delta Kappan*, *42*(7), 305-310.

Rietzschel, E. F., Nijstad, B. A., & Stroebe, W. (2007). Relative accessibility of domain knowledge and creativity: The effects of knowledge activation on the quantity and originality of generated ideas. *Journal of Experimental Social Psychology*, *43*(6), 933-946. https://doi.org/10.1016/j.jesp.2006.10.014

Runco, M. A., & Jaeger, G. (2012). The Standard Definition of Creativity. *Creativity Research Journal*, *24*, 92-96. https://doi.org/10.1080/10400419.2012.650092

Runco, M. A., & Charles, R. E. (1993). Judgments of originality and appropriateness as predictors of creativity. *Personality and Individual Differences*, *15*(5), 537-546. https://doi.org/10.1016/0191-8869(93)90337-3

Runco, M. A., & Mraz, W. (1992). Scoring Divergent Thinking Tests Using Total Ideational Output and a Creativity Index. *Educational and psychological measurement*, *52*(1), 213-221. https://doi.org/10.1177/001316449205200126

Shaw, A. (2021). It Works…but Can We Make It Easier? A Comparison of Three Subjective Scoring Indexes in the Assessment of Divergent Thinking. *Thinking Skills and Creativity*, *40*, 100789. https://doi.org/10.1016/j.tsc.2021.100789

Silvia, P., Martin, C., & Nusbaum, E. (2009). A snapshot of creativity: Evaluating a quick and simple method for assessing divergent thinking. *Thinking Skills and Creativity*, *4*, 79-85. https://doi.org/10.1016/j.tsc.2009.06.005

Silvia, P. J., Barona, C. M., Cram, J. T., Hess, K. I., Martinez, J. L., Richard, C. A., Winterstein, B. P., & Willse, J. T. (2008). Assessing Creativity With Divergent Thinking Tasks: Exploring the Reliability and Validity of New Subjective Scoring Methods. *Psychology of Aesthetics, Creativity, and the Arts*, *2*(2), 68-85. https://doi.org/10.1037/1931-3896.2.2.68

Stein, M. I. (1953). Creativity and culture. *The journal of psychology*, *36*(2), 311-322.

Sternberg, R. J., & Lubart, T. I. (1999). The concept of creativity: Prospects and paradigms. In R. J. Sternberg (Ed.), *Handbook of creativity* (pp. 3–15). Cambridge University Press.

Thomas, S. D., Hathaway, D. K., & Arheart, K. L. (1992). Face Validity. *Western Journal of Nursing Research*, *14*(1), 109-112. https://doi.org/10.1177/019394599201400111

Withagen, R., & van der Kamp, J. (2018). An ecological approach to creativity in making. *New Ideas in Psychology*, *49*, 1-6.

Zijlstra, F.R.H. & Van Doorn, L. (1985). *The construction of a scale to measure perceived effort*. Delft, The Netherlands: Department of Philosophy and Social Sciences, Delft University of Technology.

Table 1

Descriptives and Correlations

	Variable	*M (SD)*	1	2	3
1	Workload	2.64 (0.53)	-	-.08	-.04
2	Nevo Face Validity	2.95 (0.87)		-	.24**
3	Outcome-related Face Validity	3.24 (0.81)			-

Note. N = 196. **significant at = $p < .01$.

Table 2

Means and Standard Deviations Across Conditions

	Individual Idea Scoring		Snapshot Scoring	
	Combined Dimension (N = 48)	Separate Dimensions (N = 49)	Combined Dimension (N = 50)	Separate Dimension (N= 49)
Workload	2.67 (0.57)	2.58 (0.48)	2.68 (0.88)	2.62 (0.53)
Outcome-related Face Validity	3.44 (0.83)	2.94 (0.89)	3.34 (0.88)	3.24 (0.83)
Nevo Face Validity	3.06 (0.73)	3.18 (0.83)	2.78 (0.82)	2.78 (0.80)

Note. Standard deviations are given in parentheses.

Appendix A

Originality	Feasibility (N)	Ideas (Scores)
Low	Low (3)	• No smoking allowed anymore (1/1) • Food check by someone/something when you buy it (2/2) • Stop selling unhealthy food or make it more expensive: pay more tax on it (2/2)
	Middle (3)	• Healthy cooking course for young parents (2/3) • Eat everything you want until you are sick to death of it and then eat more healthily (2/3) • Require everyone to write down their eating habits and confront them with their bad eating behavior (2/3)
	High (3)	• Eat fruits and vegetables every day (1/5) • Drink fresh juice every day (1/5) • Wash hands regularly (2/5)
Middle	Low (3)	• Banning sunbeds (3/2) • Mandatory home-trainer use at work, included in collective agreement (3/2) • Supermarkets are no longer allowed to sell sweets (3/2)
	Middle (3)	• Courses for parents, so they teach health-related behaviors well to their children (3/3) • Company restaurants must comply with a healthy food label, must serve healthy food (3/3) • Reward cycling to work with a mileage allowance as it is done with car allowance (3/3)
	High (3)	• Organizing a fruit party (3/4) • Do work that is not too heavy (4/4) • Not to be stingy with regard to health-related costs, put aside a certain amount of money for it (3/4)
High	Low (4)	• Make people pay per categories of 'healthy living' (4/2) • Handing out fruits for free at the university and work (4/2) • Cover the world with foam rubber, nice and soft if you fall (5/1) • Wear a radiation-free helmet and suit (5/2)
	Middle (3)	• Change traditions, e.g.: no chocolate eggs at Easter with cheese cubes, no peppercorns at Saint Nicholas, but walnuts instead, no sweets on birthdays but something savory (4/3) • Receive text messages about your required supplies (4/3) • Develop a device that allows you to check at the end of the day whether you have received enough nutrients (4/3)
	High (3)	• Founding a new band: "The Eating Dutchmen", who, in the form of rock music, hint at what is healthy for you (4/4) • Good advertisements: E.g., ‘Fruit for a job’ – rewarding good deeds (e.g., helping elderly across a road) with tasty fruit (4/4) • Genetic modification/Genetic engineering (4/4)

Note: Total Number of Ideas: 28. In High-Low Category one more, to be able to have 7 idea sets, each made up of 4 ideas.

Appendix B

Set 1: (3x Low/Low + 1x Low/Middle)

- Stop selling unhealthy food or make it more expensive: pay more tax on it
- Food check by someone/something when you buy it
- No smoking allowed anymore
- Require everyone to write down their eating habits and confront them with their bad eating behaviour

Set 2: (2x Low/Middle + 2x Low/High)

- Eat everything you want until you are sick to death of it and then eat more healthily
- Healthy cooking course for young parents
- Wash hands regularly
- Drink fresh juice every day

Set 3: (3x Middle/Low + 1x Low/High)

- Eat fruits and vegetables every day
- Supermarkets are no longer allowed to sell sweets
- Mandatory home-trainer use at work, included in collective agreement
- Banning sunbeds

Set 4: (3x Middle/Middle + 1x Middle/High)

- Reward cycling to work with a mileage allowance as it is done with car allowance
- Company restaurants must comply with a healthy food label, must serve healthy food
- Courses for parents, so they teach health-related behaviours well to their children
- Not to be stingy with regard to health-related costs, put aside a certain amount of money for it

Set 5: (2x Middle/High + 2x High/High)

- Do work that is not too heavy
- Organizing a fruit party
- Founding a new band: "The Eating Dutchmen", who, in the form of rock music, hint at what is healthy for you
- Good advertisements: E.g., ‘Fruit for a job’ – rewarding good deeds (e.g., helping elderly across a road) with tasty fruit

Set 6: (3x High/Middle + 1x High/High)

- Genetic modification/genetic engineering
- Develop a device that allows you to check at the end of the day whether you have received enough nutrients
- Change traditions, e.g.: no chocolate eggs at Easter with cheese cubes, no peppercorns at Saint Nicholas, but walnuts instead, no sweets on birthdays but something savoury
- Receive text messages about your required supplies

Set 7: (4x High/Low)

- Cover the world with foam rubber, nice and soft if you fall
- Make people pay per categories of 'healthy living'
- Wear a radiation-free helmet and suit
- Handing out fruits for free at university and work