BPA

2/3

Bollettino di Psicologia Applicata
APPLIED PSYCHOLOGY BULLETIN
Indexed in PsycINFO® – Scopus Bibliographic Database

⬡
_____
Research

▲
_____
Experiences & Tools

CONTENTS

● Research

▲ Experiences & Tools

Differences in sensory processing in children using the AULA test: A comparative analysis of auditory and visual stimuli

Fidel Rebon-Ortiz[1], Nada Ahmed-Mahmoud[2], Zaharia-Daniel Ursu[2], Amaya Lobo[1]

[1] Giunti-Nesplora, Spain
[2] Universidad del Pais Vasco, Spain

frebon@nesplora.com

♟ ABSTRACT. L'ADHD è un disturbo che colpisce principalmente bambini e adolescenti ed è caratterizzato da sintomi importanti e duraturi di disattenzione, iperattività e impulsività. Questo studio indaga il modo in cui i bambini con sviluppo tipico (TD) rispondono agli stimoli uditivi e visivi, confrontando il loro comportamento con i bambini che presentano sospetto di disturbo da deficit di attenzione e iperattività. Per la ricerca sono stati raccolti due campioni, uno composto da 1295 partecipanti e l'altro da 378 partecipanti, di età compresa tra gli 8 e i 16 anni. L'ipotesi principale afferma che non ci sarà alcuna differenza nella risposta del partecipante agli stimoli visivi e uditivi mentre la seconda suggerisce che se la prima ipotesi non è supportata, ci saranno notevoli differenze nelle risposte del campione clinico a questi stimoli.

♟ SUMMARY. ADHD is a disorder that primarily affects children and adolescents, characterised by prominent and enduring symptoms of inattention, overactivity, and impulsivity. This study investigates how children with typical development (TD) respond to auditory and visual stimuli, comparing their behaviour to clinic-referred children who are suspected of having ADHD. Two samples were collected for this analysis, one consisting of 1,295 participants and the other consisting of 378 participants between the ages of 8 and 16. The main hypothesis states that there will be no difference in the participant's response to visual and auditory stimuli. The second hypothesis suggests that if the first hypothesis is not supported, there will be noticeable differences in the responses of the clinical sample to these stimuli. Results suggest that the first hypothesis is fulfilled for all variables except for the variables mean time for correct responses and omissions. Likewise, differences are also present in the clinical sample, confirming the second hypothesis.

Keywords: *ADHD, Attention, Children, AULA test, Auditory stimuli, Visual stimuli*

INTRODUCTION

In humans, the perception of the world, including sensory information processing, is influenced by perceptions of the environment, our emotional state, and relevant information from the surrounding environment (Carrasco, 2011; Chen et al., 2015). Processing of the stimuli received is carried out through the complex cognitive process of attention, which involves selection and focus on specific information among the multitude of stimuli around us (Gabay, Gabay, Schiff & Henik, 2019; Green, Doesburg, Ward & McDonald, 2011).

Attention can be divided into visual attention and auditory attention according to the nature of the stimulus. Visual attention focuses on the ability to enhance or process important information while inhibiting or ignoring relatively irrelevant information (Steinman & Steinman, 1998). On the other hand, auditory attention is focused on the ability to recognise relevant acoustic cues, such as speech or linguistic stimuli, and sustain that attention for an age-appropriate period of time (Andrews & Dowling, 1991; Bussing, Mason, Bell, Porter & Garvan, 2010). Both cognitive processes require specific brain mechanisms (Fiebelkorn et al., 2011), which vary according to the sex of the individual (Solberg et al., 2018). In typically developed individuals (TD), different levels of attentional performance are observed in boys and girls (Climent-Martinez & Banterla, 2011).

The most prominent psychological pathology associated with poor attention is Attention Deficit Hyperactivity Disorder (ADHD) (Schmidt & Petermann, 2009). ADHD is considered one of the most common causes of mental health problems (Hoseini, Ajilian, Moghaddam, Khademi & Saeidi, 2014) and is associated with perception, learning, memory, and executive functioning (Callahan & Terry, 2015).

ADHD manifests itself in a variety of ways, with different types of symptoms in varying levels of severity. In the case of children with ADHD, they are more likely to experience poor school performance, social isolation, and antisocial behaviour than their peers and often face significant difficulties after school (Hoseini et al., 2014).

This pathology in the school population reaches a prevalence rate of 11.4% (Willcut, 2012), causing children to experience dysfunctions that affect various activities, including academic skills in the classroom and behavioural inhibition deficits (Chiang, Chen, Lo, Tseng & Gau, 2015; Imeraj et al., 2016). Furthermore, it is often chronic, with between one third and one half of the affected persisting into adulthood.

To efficiently assess and diagnose this pathology, a virtual reality (VR) test called *Attention Kids Aula* (*AULA*) is available on the market (Iriarte et al., 2016) which has been tested in more than 1326 children, translated into more than 12 languages and with a presence in more than 24 countries around the world (Attention Kids Aula, 2021).

Its clinical report examines the significance of visual and auditory stimuli in the performance of the examinee (Climent-Martinez & Banterla, 2011). This knowledge is fundamental to the development of effective strategies that support parents, caregivers, and health professionals to manage and treat the symptoms of ADHD in these children. It may also involve environmental adjustments, such as changes in lighting or noise level in a room, and the use of specific sensory therapies to help the child regulate sensory processing, thus improving concentration and behaviour control.

To improve this study, it is essential to analyse the relationship between visual and auditory stimuli in TD children and compare it with the analysis of the prevalence of such relationships in children with suspected ADHD (clinical children). Although Lin and colleagues (Lin, Chiu, Hsieh & Wang, 2023) and Simões and colleagues (Simões, Carvalho & Schmidt, 2021) have researched auditory and visual stimuli in TD and ADHD children, they did not use the AULA test for this purpose excluding the comparison of their results to this study.

The present study aims to analyse the response of typically developed children to auditory and visual stimuli and to compare it with the behaviour of children with suspected ADHD (clinical children) based on age and sex using the AULA test.

METHOD

Participants

To carry out the study, two different samples were obtained.

The first corresponds to the sample used to perform the AULA's normative study and comprises 1,295 participants (48% female), aged from 6 to 16 years (M = 10.43, SD = 2.86). This sample may be considered representative of the

population of TD children because the prevalence of ADHD in a normal population is less than 12%.

The collaboration was proposed to schools in the Basque Country and Navarre; the schools were randomly selected based on their willingness to participate in the AULA's normative study. At the same time, informed consent was obtained from the parents of all study participants, and all students in schools between the ages of 6 and 16 were free to participate in the study. Thus, every student in the selected schools who fell within the defined age range had the same opportunity to participate in the AULA's normative study. In the end, five urban charter schools participated.

The second corresponds to the clinic data sample and comprises 378 participants (28% female), aged from 8 to 16 years ($M = 11.15$, $SD = 2.41$). This sample includes suspected ADHD children, children who are symptom-positive and who visit the clinic but whose clinical diagnosis has not been obtained for this study.

The collaboration was proposed in 108 clinics in different countries, where qualified clinic staff administered the AULA test to children aged 8-16 years with suspected ADHD. The distribution of clinics per country was as follows: AR (2), CL (3), CO (2), CR (1), EC (2), ES (75), MA (1), MX (15), PE (1), PL (1), US (4) and UY (1).

Assessment tool

The assessment tool used is a virtual reality test, AULA, which was developed to measure attention in children between 6 and 16 years of age. Its virtual setting is similar to a classroom, and the perspective places the examinee on one of the desks, facing the blackboard (see Figure 1). The head movements of the examinee are captured by the VR headset and the software updates the scene accordingly, giving the examinee the real feeling of being in the classroom. In the classroom, the examinee listens to the instructions for the tasks to be performed. To complete the task, according to specific instructions, they have to press the push button each time the presented stimulus does not appear (target stimulus), or every time the presented stimulus appears. In addition, the examinees have to face a series of distractors common in this environment, e.g., noises from the street, classmates talking/ doing other things, a knocking on the door.

Figure 1 – Screenshot of the main AULA scenario

AULA is a continuous performance test (CPT) that involves two different paradigms: an X-No paradigm, where the button has to be pressed each time the examinee does not see or hear the target stimulus, and an exercise based on an X paradigm, where the examinee has to press the button whenever they see or hear the target stimulus. Likewise, before starting the test, a usability task is carried out in which the examinee has to find some balloons and pop them to become acquainted with the test. Also, note that each paradigm (X-No and X) has a training task before starting the task.

Variables

The variables that will be included in the study are described below (see Table 1).

There is an inverse linear relationship between errors (omissions/commission) and correct answers. For this reason, only omissions and commissions are examined, since the calculation of overall performance is complementary to errors on the task.

Hypothesis

The main hypothesis is that the participants' visual stimulus exposure will be identical to the participants' auditory stimulus exposure (sample based on the AULA's normative study). The second hypothesis is that these differences will be present in the sample of children with suspected ADHD (clinical children) if differences are found between auditory and visual stimuli (the first hypothesis does not hold).

Procedure

The task and data recording procedures were performed considering the AULA administration protocol (Climent-Martínez & Banterla, 2011). This protocol consists of 3 phases of administration:

The first part, before administering the test, is used to familiarise the examinee with the equipment used. Here are the instructions to be followed by the examinee during the test. The equipment is then provided to the examinee, ensuring a comfortable position and readiness to start the test. During the test, the test administrator is required to ensure the correctness of the test and the correct collection of data. Finally, at the end of the task, the data are transferred to the computer, and the administrator of the test has the responsibility to remove the equipment from the examinee.

In terms of the technical aspect of the data collection, the VR equipment receives the test data and is connected via WIFI to a computer server that receives the data from each examinee in JSON format. Subsequently, a CSV file is generated with the variables obtained during the test.

Next, the relevant variables are selected according to the objectives of this research (see Table 1) and a descriptive analysis of the variables is carried out, as well as a hypothesis test to determine whether their distribution is normal or not, and based on these results, the statistical technique is chosen.

R version 4.2.3 (R Core Team, 2020) is used, specifically the libraries: psych (Revelle, 2024) to analyse the nature of these variables, MVN (Korkmaz, Goksuluk & Zararsiz, 2014) for testing hypotheses, VCD (Meyer, Zeileis & Hornik, 2023) for Cramer's V ratio and lubridate (Grolemund & Wickham, 2011), dplyr (Wickham, François, Henry, Müller & Vaughan, 2023), stringr (Wickham, 2022) for data handling and manipulation.

Once the statistical technique has been narrowed, the significance between the stimuli of each sample must be determined to compare the results between the two samples: the sample based on the AULA's normative study and the sample of children with suspected ADHD.

Data analysis

A descriptive analysis of the variables was carried out to provide a detailed understanding of the data. The most important characteristics of the variables are summarised (see Tables 2 and 3).

Then, an Anderson-Darling test (Anderson & Darling, 1952; Marsaglia & Marsaglia, 2004) was used to test the normality of each variable according to the rank of the age and sex set by the scale. This test is a modification of the Kolmogorov-Smirnov test, which provides a larger weight for the tails. It calculates the critical values by using a specific distribution. This has the advantage of being a sensitive test, but the disadvantage of having to calculate critical values for each distribution (see Tables 4, 5, 6 and 7).

Table 1 – Description of the variables

Variables	Description
General performance according to visual stimuli	Overall number of correct answers according to visual stimuli throughout the test. This variable is related to the general performance of the examinee throughout the test.
General performance according to auditory stimuli	Overall number of correct answers according to auditory stimuli throughout the test. This variable is related to the general performance of the examinee throughout the test.
Visual omission	Total number of visual omissions throughout the test, i.e., when the person has to press the button once the visual stimulus is presented but does not do so. This variable is indicative of the level of arousal in response to the visual target stimuli.
Auditory omission	Total number of auditory omissions throughout the test, i.e., when the person has to press the button once the auditory stimulus is presented but does not do so. This variable is indicative of the level of arousal when responding to the target auditory stimuli.
Visual commission	Total number of visual commissions throughout the test, i.e., when the person should not press the button to the presented visual stimulus and, nevertheless, presses. These errors represent an index of impulsivity or the ability to inhibit the response involved in selective attention processes.
Auditory commission	Total number of auditory commissions throughout the test, i.e., when the person should not press the button to the presented auditory stimulus and, nevertheless, presses. These errors represent an index of impulsivity or the ability to inhibit the response involved in selective attention processes.
Mean time for correct answers according to visual stimuli	This measure depicts the average time passed from the presentation of the visual target stimulus until the button pressed to respond. This measure reflects the examinee's response time.
Mean time for correct answers according to auditory stimuli	This measure depicts the average time passed from the presentation of the auditory target stimulus until the button is pressed to respond. This measure reflects the examinee's response time.
Standard deviation of time based on correct answers according to visual stimuli	Indicates the consistency of reaction time in correct answers on visual stimuli. This measure is indicative of changes in sustained attention or fatigability during the test.
Standard deviation of time based on correct answers according to auditory stimuli	Indicates the consistency of reaction time in correct answers on auditory stimuli. This measure is indicative of changes in sustained attention or fatigability during the test.

In the Anderson-Darling test, the null hypothesis is that the data follow a normal distribution. The alternative hypothesis is the lack of a normal distribution of data.

Note that according to the statistical justification of AULA, the study is adapted to the sex and age groups defined on the test scales.

A person who gets a small number of correct answers may have a smaller spread of correct answers than a person who gets a large number of correct answers and has a small number of outliers. This fact makes this variable unsuitable for this study, and therefore, the time standard deviation variable based on correct answers will be removed from the analysis.

Table 2 – Descriptive analysis of the variables based on the sample from the AULA's normative study

Variable	Mean	SD	Min.	q1	Median	q3	Max.	Skew	Kurtosis
Visual omission	17.72	21.07	0	3	8	26	93	1.49	1.35
Auditory omission	7.058	12.35	0	1	3	8	87	3.57	14.22
Visual commission	8.77	9.71	0	4	7	11	82	4.9	31.11
Auditory commission	5.99	9.09	0	2	4	7	80	5.05	31.14
Mean time for correct answers according to visual stimuli	702.97	169.03	0	586.77	674	789.37	2301	1.46	8.35
Mean time for correct answers according to auditory stimuli	1044.17	161.38	0	932.45	1039.62	1150.02	2277.74	.33	3.76
Standard deviation of time based on correct answers according to visual stimuli	263.60	122.88	0	171	239.98	330.93	867.87	.94	1.01
Standard deviation of time based on correct answers according to auditory stimuli	358.89	306.45	0	275.11	343.71	414.43	10318.68	21.41	875.13

For variables with a normal distribution, the test of equal or given proportions description was used (Wilson, 1927). For variables with a non-normal distribution, Fisher's exact test was used, which is more suited to this type of distribution (Fisher, 1922).

The test of equal or given proportions is a statistical significance test used to compare the proportions between two or more groups. The test is a comparison of the observed proportions of each category in the sample with the expected proportions, which may be either a specific set of proportions or simply the overall proportion across all categories. The null hypothesis is that the proportions in each group are equal to or a perfect match to the expected proportions, while the alternative hypothesis is that the proportions is significantly different.

If the *p*-value is less than a pre-specified significance level (= .05), the null hypothesis is rejected, and it is concluded that there is sufficient evidence to state that the proportions are significantly different.

Fisher's exact test is a non-parametric statistical test that determines whether there are significant differences between two proportions. In particular, when the sample size is small, this test is useful. The null hypothesis in Fisher's test is that the proportions are equal, indicating that the variables are independent of each other. The alternative hypothesis is that the proportions are different, indicating that the variables are dependent.

This test calculates the probability of obtaining the observed distribution of the data, as well as any more extreme distributions, given the null hypothesis. The sum of these

Table 3 – Descriptive analysis of the variables based on the sample of children with suspected ADHD

Variable	Mean	SD	Min.	q1	Median	q3	Max.	Skew	Kurtosis
Visual omission	23.38	23.80	0	4	15	36	93	1.08	.23
Auditory omission	12.38	16.61	0	2	6	15.75	87	2.31	5.66
Visual commission	11.7	13.63	0	5	8	13	93	3.44	13.79
Auditory commission	10.08	14.97	0	3	5	10	86	3.05	9.68
Mean time for correct answers according to visual stimuli	586.39	290.96	0	402.82	527.22	713.52	2477	1.53	5.64
Mean time for correct answers according to auditory stimuli	1104.2	205.55	0	986.02	1118.39	1226.15	1777	−1.06	4.87
Standard deviation of time based on correct answers according to visual stimuli	299.69	151.28	0	175.4	270.61	401.48	983.14	.73	.55
Standard deviation of time based on correct answers according to auditory stimuli	384.89	126.76	0	301.52	381.75	452.41	841.25	.32	.95

probabilities is the *p*-value associated with the test. If the *p*-value is less than a pre-specified significance level (= .05), the null hypothesis is rejected and concluded that there is sufficient evidence to state that the variables are dependent, or the proportions are different.

Because Fisher's test checks whether variables are associated, the effect size is known as the strength of the association. There are several measures of association. The most prominent are φ and Cramer's V (Cramér & Harald, 1946). The cut-off values used for their classification are as follows: .1 (small), .3 (medium), and .5 (large) based on one degree of freedom. The smaller the better, as this is intended to ensure that the two variables being measured are not related.

For hypothesis testing, the median is taken as the reference value. This is used to compare the visual stimulus with the auditory stimulus. The median is a statistic that in most cases reflects the behaviour of the population very well (Ruiz-Ruano García & López Puga, 2022). A simple division operation between visual and auditory stimuli is used to calculate the weight of the stimuli.

RESULTS

Table 8 shows results based on the sample from the AULA's normative study and Table 9 shows results based on the sample of children with suspected ADHD. Finally, Tables 10, 11, 12 and 13 summarise the results according to sample (AULA's normative study or suspected ADHD) and sex (male or female).

Table 4 – Normality of each variable based on the sample from the AULA's normative study (male): Anderson-Darling test

Variable/Year	06 (n = 90)		07 (n = 88)		08 (n = 46)		09 (n = 74)		10 (n = 64)		11 (n = 65)		12-16 (n = 236)	
	st	p	st	p	st	p	st	p	st	p	st	p	st	p
Visual omission	1.52	<.001	2.71	<.001	2.20	<.001	4.97	<.001	3.44	<.001	4.94	<.001	29.50	<.001
Auditory omission	10.33	<.001	11.29	<.001	4.85	<.001	9.94	<.001	8.41	<.001	8.39	<.001	33.65	<.001
Visual commission	5.05	<.001	1.02	.01	3.41	<.001	8.22	<.001	.70	.061*	12.77	<.001	27.88	<.001
Auditory commission	6.03	<.001	3.56	<.001	3.59	<.001	11.79	<.001	1.57	<.001	11.89	<.001	41.62	<.001
Mean time for correct answers according to visual stimuli	2.88	<.001	.81	.034	1.98	<.001	.89	.02	1.33	.002	.50	.197*	3.79	<.001
Mean time for correct answers according to auditory stimuli	.54	.154*	1.06	.008	.49	.199*	.46	.25*	2.22	<.001	.53	.161*	.76	.046
Standard deviation of time based on correct answers according to visual stimuli	.75	.004	1.04	.008	.71	.056*	.65	.08*	.48	.214*	1.04	.009	6.66	<.001
Standard deviation of time based on correct answers according to auditory stimuli	.29	.603*	5.54	<.001	.44	.273*	.61	.69*	20.78	<.001	.74	.05*	.59	.118*

Legenda. * = normality is fulfilled (p>.05); n = sample size; st = statistic; p = p-value.

Table 5 – Normality of each variable based on the sample from the AULA's normative study (female): Anderson-Darling test

Variable/Year	06 (n = 71)		07 (n = 61)		08 (n = 68)		09 (n = 65)		10 (n = 72)		11 (n = 46)		12-16 (n = 231)	
	st	p	st	p	st	p	st	p	st	p	st	p	st	p
Visual omission	13.07	.002	.74	.049	4.45	<.001	3.61	<.001	7.32	<.001	5.05	<.001	25.21	<.001
Auditory omission	8.63	<.001	3.43	<.001	3.50	<.001	8.62	<.001	13.91	<.001	5.50	<.001	29.93	<.001
Visual commission	6.22	<.001	1.59	<.001	7.98	<.001	.70	.061*	.94	.015	1.15	.004	25.2	<.001
Auditory commission	6.97	<.001	1.81	<.001	7.40	<.001	1.64	<.001	3.42	<.001	1.71	<.001	50.16	<.001
Mean time for correct answers according to visual stimuli	1.10	.006	.85	.025	1.15	.004	.62	.098*	.87	.023	.33	.496*	2.75	<.001
Mean time for correct answers according to auditory stimuli	.85	.026	.30	.565*	.25	.717*	.53	.167*	.43	.289	.60	.112*	1.01	.011
Standard deviation of time based on correct answers according to visual stimuli	.52	.176*	.66	.08*	.50	.192*	.63	.095*	.60	.111	.51	.186*	6.69	<.001
Standard deviation of time based on correct answers according to auditory stimuli	12.74	<.001	.35	.449*	.51	.187*	.26	.670*	2.09	<.001	.34	.466*	1.55	<.001

Legenda. * = normality is fulfilled (*p*>.05); *n* = sample size; *st* = statistic; *p* = *p*-value.

Table 6 – Normality of each variable based on the sample of children suspected of ADHD (male): Anderson-Darling test

Variable/Year	08 (n = 45)		09 (n = 49)		10 (n = 26)		11 (n = 30)		12-16 (n = 124)	
	st	p	st	p	st	p	st	p	st	p
Visual omission	.44	.278*	1.24	.002	4.63	<.001	1.97	<.001	10.92	<.001
Auditory omission	2.94	<.001	3.88	<.001	4.55	<.001	3.39	<.001	15.12	<.001
Visual commission	5.50	<.001	1.53	<.001	7.41	<.001	3.06	<.001	12.04	<.001
Auditory commission	5.39	<.001	3.99	<.001	8.75	<.001	4.33	<.001	19.37	<.001
Mean time for correct answers according to visual stimuli	.57	.125*	2.41	<.001	.34	.462*	1.19	.003	6.51	<.001
Mean time for correct answers according to auditory stimuli	1.76	<.001	1.31	.001	.35	.441*	.88	.020	.58	.127*
Standard deviation of time based on correct answers according to visual stimuli	.19	.888*	.38	.384*	.24	.723*	.79	.034	4.76	<.001
Standard deviation of time based on correct answers according to auditory stimuli	.56	.131*	.41	.318*	.63	.081*	.32	.512*	.48	.228*

Legenda. * = normality is fulfilled (*p*>.05); *n* = sample size; *st* = statistic; *p* = *p*-value.

Table 7 – Normality of each variable based on the sample of children suspected of ADHD (female): Anderson-Darling test

Variable/Year	08 (n = 21)		09 (n = 17)		10 (n = 10)		11 (n = 11)		12-16 (n = 45)	
	st	p	st	p	st	p	st	p	st	p
Visual omission	.19	.866*	.60	.099*	.56	.104*	.63	.07*	4.63	<.01
Auditory omission	.43	.272*	.85	.022	.31	.479*	1.66	<.001	4.55	<.01
Visual commission	2.95	<.001	1.59	<.001	.74	.034	.46	.202*	7.41	<.01
Auditory commission	2.61	<.001	.79	.032	.97	.008	.53	.130*	8.75	<.01
Mean time for correct answers according to visual stimuli	.46	.229*	.34	.450*	.87	.015	.42	.265*	2.26	<.01
Mean time for correct answers according to auditory stimuli	1.04	.007	.18	.900*	.28	.552*	.30	.512*	.30	.565*
Standard deviation of time based on correct answers according to visual stimuli	.28	.603*	.41	.303*	.30	.509*	.37	.341*	.95	.014
Standard deviation of time based on correct answers according to auditory stimuli	.41	.309*	.44	.247*	.22	.741*	.24	.676*	.28	.599*

Legenda. * = normality is fulfilled (p>.05); n = sample size; st = statistic; p = p-value.

Table 8 – Significance and prevalence of each variable by sex and age from the sample of the AULA's normative study

Group	Variable	Sig.	Weight	Cramer	Prev
Male					
06 (*n* = 90)					
	Mean time for correct answers	yes	1342	.149	auditory
	Omission	yes	203	.250	visual
	Commission	no	889	.087	none
07 (*n* = 88)					
	Mean time for correct answers	yes	1369	.038	auditory
	Omission	yes	179	.242	visual
	Commission	no	.75	.027	none
08 (*n* = 46)					
	Mean time for correct answers	yes	1451	.082	auditory
	Omission	no	276	.113	none
	Commission	no	.55	.030	none
09 (*n* = 74)					
	Mean time for correct answers	no	1529	.019	none
	Omission	no	278	.141	none
	Commission	no	455	.097	none
10 (*n* = 64)					
	Mean time for correct answers	yes	1575	.100	auditory
	Omission	no	353	.167	none
	Commission	no	444	.089	none
11 (*n* = 65)					
	Mean time for correct answers	yes	1.64	.036	auditory
	Omission	no	286	.141	none
	Commission	no	.5	.072	none
12-16 (*n* = 236)					
	Mean time for correct answers	yes	1662	.047	auditory
	Omission	no	667	.023	none
	Commission	no	429	.087	none

continued

Group	Variable	Sig.	Weight	Cramer	Prev
Female					
06 (*n* = 71)					
	Mean time for correct answers	yes	1265	.063	auditory
	Omission	yes	216	.225	visual
	Commission	no	857	.011	none
07 (*n* = 61)					
	Mean time for correct answers	yes	1.46	.045	auditory
	Omission	yes	208	.191	visual
	Commission	no	.75	.051	none
08 (*n* = 68)					
	Mean time for correct answers	yes	1465	.030	auditory
	Omission	no	.32	.005	none
	Commission	no	.8	.023	none
09 (*n* = 65)					
	Mean time for correct answers	no	1588	.004	none
	Omission	no	273	.131	none
	Commission	no	.5	.078	none
10 (*n* = 72)					
	Mean time for correct answers	no	1584	.019	none
	Omission	no	333	.119	none
	Commission	no	333	.196	none
11 (*n* = 46)					
	Mean time for correct answers	no	1557	.027	none
	Omission	no	.2	.064	none
	Commission	no	.5	.057	none
12-16 (*n* = 231)					
	Mean time for correct answers	yes	1604	.071	auditory
	Omission	no	333	.064	none
	Commission	no	.4	.075	none

Table 9 – Significance and prevalence of each variable by sex and age from the sample of children suspected of ADHD

Group	Variable	Sig	Weight	Cramer	Prev
Male					
08 (*n* = 45)					
	Mean time for correct answers	yes	1725	.117	auditory
	Omission	yes	.4	.155	visual
	Commission	no	818	.018	none
09 (*n* = 49)					
	Mean time for correct answers	yes	1.93	.206	auditory
	Omission	yes	321	.181	visual
	Commission	no	727	.100	none
10 (*n* = 26)					
	Mean time for correct answers	yes	1975	.101	auditory
	Omission	no	765	.011	none
	Commission	no	765	.040	none
11 (*n* = 30)					
	Mean time for correct answers	yes	2196	.119	auditory
	Omission	yes	.3	.153	visual
	Commission	no	857	.071	none
12-16 (*n* = 124)					
	Mean time for correct answers	yes	2379	.284	auditory
	Omission	no	.5	.038	none
	Commission	no	429	.108	none
Female					
08 (*n* = 21)					
	Mean time for correct answers	yes	1719	.092	auditory
	Omission	no	436	.066	none
	Commission	no	1111	.035	none
09 (*n* = 17)					
	Mean time for correct answers	yes	1663	.058	auditory
	Omission	yes	.25	.225	visual

continued

Group	Variable	Sig	Weight	Cramer	Prev
	Commission	no	1	.111	none
10 (*n* = 10)					
	Mean time for correct answers	yes	1797	.130	auditory
	Omission	no	.5	.008	none
	Commission	no	889	.080	none
11 (*n* = 11)					
	Mean time for correct answers	yes	2302	.171	auditory
	Omission	no	438	.140	none
	Commission	no	1.75	.106	none
12-16 (*n* = 45)					
	Mean time for correct answers	yes	2687	.178	auditory
	Omission	no	.5	.052	none
	Commission	no	.5	.066	none

Table 10 – Summary of significance on the sample of the AULA's normative study (male)

Variable/Year	06 (*n* = 90)	07 (*n* = 88)	08 (*n* = 46)	10 (*n* = 64)	11 (*n* = 65)	12-16 (*n* = 236)
Mean time for correct answers	auditory	auditory	auditory	auditory	auditory	auditory
Omission	visual	visual				

Table 11 – Summary of significance on the sample of the AULA's normative study (female)

Variable/Year	06 (*n* = 71)	07 (*n* = 61)	08 (*n* = 68)	12-16 (*n* = 231)
Mean time for correct answers	auditory	auditory	auditory	auditory
Omission	visual	visual		

Table 12 – Summary of significance on the sample of children suspected of ADHD (male)

Variable/Year	08 (n = 45)	09 (n = 49)	10 (n = 26)	11 (n = 30)	12-16 (n = 124)
Mean time for correct answers	auditory	auditory	auditory	auditory	auditory
Omission	visual	visual		visual	

Table 13 – Summary of significance on the sample of children suspected of ADHD (female)

Variable/Year	08 (n = 21)	09 (n = 17)	10 (n = 10)	11 (n = 11)	12-16 (n = 124)
Mean time for correct answers	auditory	auditory	auditory	auditory	auditory
Omission		visual			

DISCUSSION

Visual and auditory stimuli can interact with each other to improve understanding of the environment. For example, in a video game, hearing helps the individual understand what is being played. Similarly, sight can help the person understand the rhythm and movement of the notes on the sheet music when listening to music.

Prioritising one type of stimulus over the other can help children regulate sensory processing, improve concentration, and foster better behavioural control. When a significant value is observed in these variables, it may signify challenges in the child's interpretation of this stimulus compared to the other. This is due to the unique nature of these variables, where lower values indicate better performance, different from the general performance variable.

In the first phase of the study, to answer the first hypotheses presented, the variables were analysed to identify the type of stimulus that was more significant in the population represented by the sample from the AULA's normative study. In this study, two variables were found to be significant among stimuli: the mean time for correct answers and the omission. In conclusion, the auditory stimulus was predominant for the first variable and the visual stimulus was predominant for the second variable. However, this significance for the visual stimulus occurs only up to the age of 9 years, and no distinction is made between boys and girls. For the remaining variables, no significance was found between auditory and visual stimuli.

A positive outcome is the absence of significant differences between visual and auditory stimuli in a wide range of variables. This indicates that any noteworthy findings in test performance, visual or auditory, should be emphasised by clinicians given the lack of overall significance in the population.

These findings are instrumental in establishing a knowledge base on the importance of visual and auditory stimuli as the primary variables in AULA. When a test

is administered to an examinee and their performance deviates from that of the general population, this incongruity must be explicitly noted in the report to the clinician, as this information is crucial to helping children regulate their sensory processing, thus improving their ability to concentrate and control their behaviours.

In the second phase of the study, analysing the sample of children suspected of ADHD, it was observed that the mean time for correct answers variable continued to show the same behaviour as in the first phase of the study. That is, the auditory stimuli have shown the greatest significance. On the contrary, a greater diversity of results was found for the omission variable. Children in the AULA normative sample show a significant preference for visual stimuli between the ages of 6 and 7, but after this age, there is no clear preference for either stimulus. In contrast, in both boys and girls in the sample of children suspected of having ADHD, the preference for visual stimuli persists in boys until the age of 11, while in girls this preference is reduced at the age of 9.

CONCLUSIONS AND FUTURE PERSPECTIVES

The main hypothesis, that the participants' visual stimulus exposure will be identical to the participants' auditory stimulus exposure (sample based on the normative AULA study), is fulfilled for all variables except for the following: mean time for correct responses and omissions. There is a higher sensitivity toward auditory stimuli in the meantime for correct responses variable and a higher sensitivity towards visual stimuli in the omission variable, but this variable is no longer significant for either sex from the age of seven years.

There are also differences in the sample of children with suspected ADHD, which confirms the second hypothesis. It is important to note that the significance of visual stimulus in the omission variable extends until the age of 11 for boys, while for girls, it persists until the age of 9. For this reason, the omission variable shows a different trend between boys and girls with suspected ADHD.

In general, AULA variables show no significance between visual and auditory stimuli, except for two of them (mean time for correct responses and omissions). In addition, it is important to note that these variables work inversely to the general performance variables, since a shorter mean time for correct responses is better than a longer mean time for correct responses. In addition, a small number of omissions is more appropriate than a large number of omissions. This is an aspect to be considered when drawing conclusions from the results obtained.

The results of the present study are limited by the low number of cases in some age and sex subgroups. Therefore, it is recommended that the sample be expanded and include two new study groups: one exclusively with children with ADHD and the other with TD children. This extension would enrich the conclusions of this study.

References

ANDERSON, T.W., & DARLING, D.A. (1952). Asymptotic theory of certain "goodness of fit" criteria based on stochastic processes. *The Annals of Mathematical Statistics*, *23* (2), 193-212. doi.org/10.1214/aoms/1177729437

ANDREWS, M.W., & DOWLING, W.J. (1991). The development of perception of interleaved melodies and control of auditory attention. *Music Perception*, *8* (4), 349-368. doi.org/10.2307/40285518

Attention Kids Aula (2021, September 9). Nesplora. https://nesplora.com/producto/attention-kids-aula/

BUSSING, R., MASON, D.M., BELL, L., PORTER, P., & GARVAN, C. (2010). Adolescent outcomes of childhood attention-deficit/hyperactivity disorder in a diverse community sample. *Journal of the American Academy of Child and Adolescent Psychiatry*, *49* (6), 595-605. doi.org/10.1097/00004583-201006000-00008

CALLAHAN, P.M., & TERRY, A.V. (2015). *Attention*. In K. Kantak

& J. Wettstein (Eds.), *Cognitive enhancement. Handbook of experimental pharmacology, vol 228*. Cham, CH: Springer. doi. org/10.1007/978-3-319-16522-6_5

CARRASCO, M. (2011). Visual attention: The past 25 years. *Vision Research*, *51* (13), 1484-1525. doi.org/10.1016/j.visres.2011.04.012

CHEN, T., MICHELS, L., SUPEKAR, K., KOCHALKA, J., RYALI, S., & MENON, V. (2015). Role of the anterior insular cortex in integrative causal signaling during multisensory auditory-visual attention. *The European Journal of Neuroscience*, *41* (2), 264-274. doi.org/10.1111/ejn.12764

CHIANG, H.-L., CHEN, Y.-J., LO, Y.-C., TSENG, W.-Y. I., & GAU, S.S. (2015). Altered white matter tract property related to impaired focused attention, sustained attention, cognitive impulsivity and vigilance in attention-deficit/ hyperactivity disorder. *Journal of Psychiatry & Neuroscience: JPN*, *40* (5), 325-335. doi.org/10.1503/jpn.140106

CLIMENT-MARTÍNEZ, G., & BANTERLA, F. (2011). *AULA. Manual teórico*. San Sebastián: Nesplora.

CRAMÉR, H. (1946). A contribution to the theory of statistical estimation. *Scandinavian Actuarial Journal*, *1946* (1), 85-94. doi. org/10.1080/03461238.1946.10419631

FIEBELKORN, I.C., FOXE, J.J., BUTLER, J.S., MERCIER, M.R., SNYDER, A.C., & MOLHOLM, S. (2011). Ready, set, reset: Stimulus-locked periodicity in behavioral performance demonstrates the consequences of cross-sensory phase reset. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *31* (27), 9971-9981. doi.org/10.1523/JNEUROSCI.1338-11.2011

FISHER, R.A. (1922). On the interpretation of χ^2 from contingency tables, and the calculation of P. *Journal of the Royal Statistical Society*, *85* (1), 87. doi.org/10.2307/2340521

GABAY, Y., GABAY, S., SCHIFF, R., & HENIK, A. (2019). Visual and auditory interference control of attention in developmental dyslexia. *Journal of the International Neuropsychological Society: JINS*, *26* (4), 407-417. doi.org/10.1017/S135561771900122X

GOMES, H., WOLFSON, V., & HALPERIN, J.M. (2007). Is there a selective relationship between language functioning and auditory attention in children? *Journal of Clinical and Experimental Neuropsychology*, *29* (6), 660-668. doi. org/10.1080/13803390600920455

GREEN, J.J., DOESBURG, S.M., WARD, L.M., & McDONALD, J.J. (2011). Electrical neuroimaging of voluntary audiospatial attention: Evidence for a supramodal attention control network. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *31* (10), 3560-3564. doi.org/10.1523/JNEUROSCI.5758-10.2011

GROLEMUND, G., & WICKHAM, H. (2011). Dates and times made easy with lubridate. *Journal of Statistical Software*, *40* (3). doi. org/10.18637/jss.v040.i03

HOSEINI, B., AJILIAN, M., MOGHADDAM, H., KHADEMI, G., & SAEIDI, M. (2014). Attention Deficit Hyperactivity Disorder (ADHD) in children: A short review and literature. *International Journal of Pediatrics*, *2*, 443-450. doi.org/10.22038/ijp.2014.3749

IMERAJ, L., ANTROP, I., ROEYERS, H., DEBOUTTE, D., DESCHEPPER, E., BAL, S., & SONUGA-BARKE, E. (2016). The impact of idle time in the classroom: Differential effects on children with ADHD. *Journal of Attention Disorders*, *20* (1), 71-81. doi.org/10.1177/1087054713478464

IRIARTE, Y., DIAZ-ORUETA, U., CUETO, E., IRAZUSTABARRENA, P., BANTERLA, F., & CLIMENT, G. (2016). AULA-advanced virtual reality tool for the assessment of attention: Normative study in Spain. *Journal of Attention Disorders*, *20* (6), 542-568. doi.org/10.1177/1087054712465335

KORKMAZ, S., GOKSULUK, D., & ZARARSIZ, G. (2014). MVN: An R package for assessing multivariate normality. *The R Journal*, *6* (2), 151. doi.org/10.32614/rj-2014-031

LIN, H.-Y., CHIU, E.-C., HSIEH, H.-C., & WANG, P.-J. (2023). Gender differences in auditory and visual attentional performance in children with and without ADHD. *Archives of Clinical Neuropsychology: The Official Journal of the National Academy of Neuropsychologists*, *38* (6), 891-903. doi.org/10.1093/arclin/acad019

LIN, H.-Y., HSIEH, H.-C., LEE, P., HONG, F.-Y., CHANG, W.-D., & LIU, K.-C. (2017). Auditory and visual attention performance in children with ADHD: The attentional deficiency of ADHD is modality specific. *Journal of Attention Disorders*, *21* (10), 856-864. doi.org/10.1177/1087054714542004

MARSAGLIA, G., & MARSAGLIA, J. (2004). Evaluating the Anderson-darling distribution. *Journal of Statistical Software*, *9* (2). doi.org/10.18637/jss.v009.i02

MEYER, D., ZEILEIS A., & HORNIK, K. (2023). *Visualizing categorical data [R package vcd version 1.4-12]*. https://CRAN.R-project.org/package=vcd

R CORE TEAM (2020). *The R project for statistical computing*. R-project.org. https://www.R-project.org/

REVELLE, W. (2024). *Procedures for psychological, psychometric, and personality research [R package psych version 2.4.1]*. https://CRAN.R-project.org/package=psych

RUIZ-RUANO GARCÍA, A.M., & LÓPEZ PUGA, J. (2022). *Análisis de datos en psicología y educación*. Aula Magna Proyecto Clave, McGraw Hill.

RUSSELL, E.L. (2015). *Auditory and visual sustained attention on*

tasks with varied motivation and cognitive loads in children with and without ADHD. https://thescholarship.ecu.edu/handle/10342/4993

SCHMIDT, S., & PETERMANN, F. (2009). Developmental psychopathology: Attention Deficit Hyperactivity Disorder (ADHD). *BMC Psychiatry*, *9* (1), 58. doi.org/10.1186/1471-244X-9-58

SIMÕES, E.N., CARVALHO, A.L.N., & SCHMIDT, S.L. (2021). The role of visual and auditory stimuli in continuous performance tests: Differential effects on children with ADHD. *Journal of Attention Disorders*, *25* (1), 53-62. doi.org/10.1177/1087054718769149

SOLBERG, B.S., HALMØY, A., ENGELAND, A., IGLAND, J., HAAVIK, J., & KLUNGSØYR, K. (2018). Gender differences in psychiatric comorbidity: A population-based study of 40 000 adults with attention deficit hyperactivity disorder. *Acta Psychiatrica Scandinavica*, *137* (3), 176-186. doi.org/10.1111/acps.12845

STEINMAN, S.B., & STEINMAN, B.A. (1998). Vision and attention. I: Current models of visual attention. *Optometry and Vision Science: Official Publication of the American Academy of Optometry*, *75* (2), 146-155. doi.org/10.1097/00006324-199802000-00018

WICKHAM, H. (2022). *Simple, consistent wrappers for common string operations.* https://stringr.tidyverse.org, https://github.com/tidyverse/stringr

WICKHAM, H., FRANÇOIS, R., HENRY, L., MÜLLER, K., & VAUGHAN, D. (2023). *dplyr: A grammar of data manipulation.* Independently Published.

WILLCUTT, E.G. (2012). The prevalence of DSM-IV attention-deficit/hyperactivity disorder: A meta-analytic review. *Neurotherapeutics: The Journal of the American Society for Experimental NeuroTherapeutics*, *9* (3), 490-499. doi.org/10.1007/s13311-012-0135-8

WILSON, E.B. (1927). Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, *22* (158), 209. doi.org/10.2307/2276774

Development and validation of Tripartite Schadenfreude Scale

Sana Safia[1], Adnan Adil[2], Saba Ghayas[3], Sultan Shujja[3]

[1] *Department of Psychology, University of Sargodha, Sargodha, Pakistan*
[2] *Corresponding Author, PhD, Professor, Department of Psychology, Government College Women University Sialkot, Pakistan*
[3] *PhD, Assistant Professor, Department of Psychology, University of Sargodha, Sargodha, Pakistan*

livespirit786@yahoo.com

✿ ABSTRACT. Questo studio mirava a sviluppare una misura self-report della schadenfreude basata su un modello motivazionale tripartito. Nello Studio 1, è stato sviluppato un insieme di 36 item: l'analisi fattoriale di conferma (CFA) degli elementi della TSS in un campione intenzionale di 300 studenti dell'Università di Sargodha ha rivelato una soluzione a tre fattori di secondo ordine coerente con il modello motivazionale tripartito di schadenfreude. Lo Studio 2 ha replicato la struttura fattoriale della TSS in un campione intenzionale di 219 studenti anch'essi dell'Università di Sargodha. La TSS non era correlata con la vergogna o il senso di colpa, indicando così una validità discriminante. Nel complesso, dunque, questi risultati suggeriscono che la TSS di nuova concezione è una misura psicometricamente valida della schadenfreude tripartita.

✿ SUMMARY. *Entrenched in the tripartite motivational model of schadenfreude (Wang, Lilienfeld & Rochat, 2019), this study developed and validated a self-report measure of schadenfreude. In Study 1, a pool of 36 items was developed after a thorough review of the literature and expert opinions. The content validity index for the items and scale was established after obtaining the ratings of four experts. Confirmatory factor analysis (CFA) of the item pool of the Tripartite Schadenfreude Scale (TSS) in a purposive sample of 300 students at the University of Sargodha revealed a second-order three-factor solution consistent with the tripartite motivational model of schadenfreude. The first-order factor loadings ranged from .60 to .80 and their Cronbach's alpha coefficients of reliability ranged from .73 to .94. TSS and its three factors demonstrated a strong positive correlation with the Perceived Schadenfreude Scale (Batool, 2014), which established evidence of convergent validity. Study 2 replicated the factorial structure of TSS in a purposive sample of 219 students at the University of Sargodha. Moreover, TSS did not correlate with shame or guilt, indicating discriminant validity. Overall, these findings suggest that the newly developed TSS is a psychometrically sound measure of tripartite schadenfreude.*

Keywords: *Tripartite schadenfreude, Aggression schadenfreude, Rivalry schadenfreude, Justice schadenfreude*

INTRODUCTION

According to van Dijk and colleagues (van Dijk, Wesseling, Ouwerkerk & van Koningsbruggen, 2010), experiencing pleasure from the misfortune of others is a commonly observed phenomenon. Although individuals tend to disapprove of such feelings when explicitly asked, there are instances in which a smile emerges involuntarily when witnessing another person's misfortune. This phenomenon, known as schadenfreude, is intricately connected to how individuals interpret the misfortune of others, and various factors influence the degree of pleasure experienced. These factors include the severity of the misfortune, the identity of the victim, and the individual's personal experiences with the victim (Chen & Lee, 2020; Thompson & Martinez, 2023). According to Heider (1958), the German word schadenfreude means harm joy and it refers to the joy or pleasure that people feel on the misfortune of others. A person who feels pleasure in other's misfortunes is termed schadenfroh. The current research is based on the tripartite motivational model of schadenfreude, based on a review study by Wang et al. (2019).

Tripartite motivational model of schadenfreude

The tripartite motivational model of schadenfreude (Wang et al., 2019) examines the underlying motivations behind the experience of schadenfreude. According to this model, schadenfreude is driven by three primary motivational components: aggression, rivalry and justice.

The aggression component of schadenfreude suggests that individuals may experience pleasure when witnessing the misfortune of others, because it satisfies their aggressive tendencies. This may be particularly relevant when a person experiencing misfortune is perceived as causing harm or injustice to others. The rivalry component focuses on the comparison between oneself and a person experiencing a misfortune. Individuals may experience schadenfreude if they perceive the misfortune of others as enhancing their relative standing or superiority. It can be driven by feelings of envy, rivalry, or desire to maintain a sense of superiority. The justice component highlights the roles of perceived fairness and justice in schadenfreude. Individuals may derive pleasure from seeing the misfortune of others if they believe it is a form of poetic justice or retribution for the perceived wrongdoing of the person experiencing misfortune.

The present research aimed to develop an indigenous measure of schadenfreude based on a tripartite motivational model (Wang et al., 2019) for Pakistani population. For this purpose, the standard procedure for scale construction is employed. The following section briefly describes each tripartite motivational model component.

– *Aggression schadenfreude.* According to Wang et al. (2019), aggression schadenfreude is a specific form of schadenfreude that revolves around deriving pleasure and satisfaction from witnessing the misfortunes or failures of others, particularly those perceived as competitors or adversaries. It is characterized by feelings of aggression and hostility towards the target of schadenfreude. Unlike other forms of schadenfreude, such as justice-based schadenfreude, aggression schadenfreude is primarily motivated by hostile intentions. In aggression schadenfreude, individuals experience a sense of pleasure and vindication when their rivals or enemies encounter setbacks, failures, or suffering. The target's misfortune is seen as a form of triumph or validation of one's superiority or dominance. The pleasure derived from aggression schadenfreude may stem from a desire to see the downfall of perceived threats or competitors, providing a sense of satisfaction and a boost to one's self-esteem.

– *Rivalry schadenfreude.* According to Wang et al. (2019), rivalry schadenfreude refers to the experience of pleasure or satisfaction derived from the misfortune or failure of individuals with whom one shares a competitive relationship or rivalry. It involves taking joy in the setbacks or difficulties encountered by rivals or competitors. Rivalry schadenfreude is often driven by the desire for personal success or achievement relative to others. When individuals perceive their rivals as threats or obstacles to their goals or aspirations, witnessing their failures can elicit a sense of pleasure or satisfaction. This can provide a sense of validation and superiority, as the misfortune of rivals is seen as a confirmation of one's relative success or competence (van Dijk & Ouwerkerk, 2014).

This form of schadenfreude can manifest in various domains including academics, sports, professional settings, and personal relationships. For example, in sports, fans may experience rivalry schadenfreude when their team's rival loses a game or faces defeat (Demir & Rigoni, 2017). In academic or professional settings, individuals may enjoy seeing that their competitors fail to achieve

their desired outcomes or face setbacks. Van de Ven et al. (2015) found that individuals who harbor malicious envy toward someone tend to experience heightened levels of schadenfreude when that person encounters a misfortune.

–   *Justice schadenfreude.* Wang et al. (2019) delineated justice schadenfreude as the experience of pleasure or satisfaction that arises when witnessing the misfortunes or suffering of others who are perceived as deserving of punishment or experiencing a consequence for their actions. It is rooted in a sense of fairness, in which individuals believe that the harm suffered by the target of schadenfreude is justified based on their past behaviors or transgressions. The focus in justice schadenfreude is on the perceived fairness and appropriateness of punishment rather than deriving pleasure from suffering itself.

    According to Smith and van Dijk (2018), the concept of justice schadenfreude is closely tied to the notion of retributive justice, which emphasizes that individuals should face consequences proportional to their actions. When individuals perceive that justice is served, they may experience a positive emotional response, including schadenfreude. This form of schadenfreude is driven by the belief that the target's misfortune is deserved, and serves as a form of moral vindication. Piskorz and Piskorz (2009) further established a positive association between schadenfreude and the perceived deservingness of the victim, finding that greater pleasure was derived when misfortune was considered justified or deserved.

    Justice schadenfreude can be observed in various contexts, such as witnessing the downfall of individuals who have committed crimes, experienced professional or personal failures due to unethical behavior, or faced consequences for their harmful actions. It can also be directed towards groups or institutions that are seen as deserving of punishment or retribution. Overall, justice schadenfreude reflects a complex emotional response to the perceived alignment of a person's misfortune with their past actions rooted in the desire for fairness and moral accountability (Wang et al., 2019).

–   *Developmental perspective of schadenfreude.* The developmental trajectory proposed by the tripartite motivation model of schadenfreude (van Dijk et al., 2019) suggests that different forms of schadenfreude emerge at different stages of development and are influenced by individuals evolving cognitive and social capacities.

    According to van Dijk and colleagues (2019), aggression schadenfreude tends to emerge early in development. In childhood, children may engage in aggressive behaviors toward their peers or siblings. When they witness the misfortune or suffering of others, particularly those they perceive as competitors or adversaries, they may experience a sense of pleasure or satisfaction. This aggression schadenfreude is rooted in feelings of dominance, power, and gratification derived from seeing others in a subordinate position. Tajfel and Turner (1986) stated that infants at least by 9-14 months preferred individuals who harm dissimilar others to those who help them, an effect more pronounced in older infants. This finding suggests that infants' social evaluations are governed by a rudimentary sense of social identity rooted in similarity/dissimilarity judgments. Cikara and colleagues (Cikara, Botvinick & Fiske, 2011) argued that perceiving others as dissimilar motivates negative evaluation which provokes aggression. Hamlin and colleagues (Hamlin, Mahajan, Liberman & Wynn, 2013) stated that dissimilar others are previewed as out-group members, and their punishment is perceived as deserving and enjoyable. Therefore, infants' positive evaluation of harming dissimilar others may be a harbinger of aggression schadenfreude that derives from a concern for social identity.

As individuals progress through adolescence and adulthood, social comparisons and competition become more pronounced. Rivalry schadenfreude arises from a desire to outperform or gain an advantage over others. When individuals witness failures, setbacks, or misfortunes of their rivals or competitors, they may experience pleasure or satisfaction. This form of schadenfreude is driven by the need for personal success and validation obtained from seeing others who are perceived as threats or obstacles encountering difficulties (van Dijk et al., 2019). Research suggests that rivalry schadenfreude originates from concerns regarding social comparison. Steinbeis and Singer (2013) examined envy and schadenfreude in 7-13-year-old children in a reward and punishment task. They found that even children experience schadenfreude and envy toward their rivals. The children competed with an anonymous child to win a prize. In another study, Shamay and colleagues (2014) showed that 24-month-old infants who previously expressed jealousy toward another infant exhibited behavior indicative of schadenfreude when the rival infant lost his/her favorable position.

With further cognitive and moral development,

individuals begin to understand and appreciate social norms, fairness, and moral accountability. Justice-based schadenfreude arises when individuals perceive someone as deserving punishment or experiencing consequences for their actions. Witnessing the misfortune or suffering of these individuals elicits pleasure or satisfaction based on the belief that justice is being served. This form of schadenfreude is rooted in a sense of moral righteousness and the alignment of the target's misfortune with past behaviors (van Dijk et al., 2019). According to Nobes and colleagues (Nobes, Panagiotaki & Pawson, 2009), developmental research revealed that schadenfreude may trace its roots partly to a norm-based moral evaluation in children already evident in preschool years. Hamlin and Wynn (2011) argued that concern for social justice may even trace its developmental roots to social evaluation demonstrated in few-month-old infants. Three-month-old infants prefer puppets who help rather than hinder one another. This early emerging social discrimination is not only crucial for navigating the social world but may also be a developmental precursor to schadenfreude.

The developmental trajectory of schadenfreude suggests that aggression schadenfreude serves as an initial foundation, followed by the emergence of rivalry and justice-based schadenfreude as individuals mature and develop their cognitive and moral capacities. This trajectory highlights the interplay between emotional responses, social dynamics, and moral reasoning throughout different stages of development.

Measurement of tripartite schadenfreude

Schadenfreude has important implications in social and interpersonal relationships. It is conceived as a negative emotional state, as it is incongruent with the victim's emotional state. For instance, in cases of bullying or harassment, some individuals find pleasure in witnessing their victims' suffering or misfortune. Likewise, rather than empathizing with someone's success or accomplishments, certain individuals may enjoy observing their failures or setbacks. Instead of celebrating collective achievements, some individuals derive satisfaction from competitors' misfortunes. Gossip frequently involves spreading negative information about others, and certain individuals derive enjoyment from hearing about the misfortune or scandals surrounding people they are acquainted with. Smith and colleagues (Smith, Powell, Combs & Schurtz, 2009) reason that while schadenfreude is commonly associated with negative social connotations, it is important to acknowledge that it can arise as a natural and instinctive emotional response in humans.

The tripartite motivational model of schadenfreude is considered superior to its unidimensional conception for several reasons. First, the tripartite motivational model provides a more comprehensive explanation of schadenfreude by recognizing that it consists of multiple sub-forms with different underlying motivations. It acknowledges that schadenfreude can arise from various psychological processes and goals such as self-enhancement, social comparison, and justice-related concerns. In contrast, the unidimensional conception treats schadenfreude as a single, uniform phenomenon, without accounting for these diverse motivations.

Second, the tripartite model distinguishes between sub-forms of schadenfreude, namely, rivalry, aggression, and justice schadenfreude. This differentiation allows a clearer understanding of the underlying psychological mechanisms and contextual factors that contribute to each type of schadenfreude. By contrast, the unidimensional conception does not differentiate between these sub-forms, leading to a less nuanced understanding of the phenomenon.

Third, by considering the distinct motivations and processes involved in different sub-forms of schadenfreude, the tripartite model has greater predictive power. This can account for variations in schadenfreude experiences across different situations and individuals. In contrast, unidimensional conceptions may struggle to explain the heterogeneity observed in schadenfreude responses.

Fourth, the tripartite model's comprehensive understanding of schadenfreude has practical implications for various domains. For example, it can inform interventions aimed at reducing harmful or malicious expressions of schadenfreude, while promoting prosocial behaviors. A nuanced understanding of the different motivations behind schadenfreude can also aid in designing effective communication strategies and conflict-resolution techniques.

Overall, the tripartite motivational model of schadenfreude offers a more comprehensive, differentiated, and predictive framework than the unidimensional model. Its multidimensional approach recognizes the complexity and

diversity of schadenfreude experiences, leading to a deeper understanding of the phenomenon and its implications.

Owing to the aforementioned characteristics of the tripartite motivational model of schadenfreude, this study aims to develop a psychometrically sound measure of schadenfreude in consonance with the tripartite motivational model of schadenfreude (Wang et al., 2019). The current research constructed and psychometrically tested a measure that operationalizes schadenfreude in terms of rivalry, aggression, and justice schadenfreude; thus, it has yielded empirical support for the tripartite motivational model of schadenfreude. The existing measures of schadenfreude do not espouse a sound theoretical framework, and they do not identify the various dimensions of schadenfreude.

To the best of our knowledge, there are three scales for measuring schadenfreude. The first was indigenously developed by Batool (2014) and conceptualized schadenfreude in terms of the superiority theory of humor (Hurley et al., 2011). This theory explains schadenfreude in terms of humor, people's experience of the misfortune of others, and feeling superior. This scale was unidimensional, and it only measured a single aspect of schadenfreude in terms of the humor that people feel at the misfortune of others. The second scale is van Dijk's *Schadenfreude and Sympathy Scale* (van Dijk, Ouwerkerk, Goslinga & Nieweg, 2005), which is a short measure of schadenfreude (5 items) and sympathy (3 items). This scale is also unidimensional and conceives of schadenfreude in very general terms of humor on the misfortunes of others. The third measure is the *Trait Schadenfreude Scale* (Baren, 2017), which was developed to measure individual differences in trait schadenfreude to understand the degree to which people differ in their degree of joy they might feel when learning others' downfall. This scale is also unidimensional and conceives of schadenfreude in terms of humor or fun that people may feel about the everyday misfortune of others. Thus, it becomes clear that the available measures of schadenfreude are not based on any exclusive theory of schadenfreude, and all of them have conceived of schadenfreude as a unidimensional construct in terms of fun, amusement, or joy that individuals experience in the suffering of others in everyday situations.

To the best of our knowledge, schadenfreude sub-forms are yet to be operationally defined following the tripartite motivational model through a psychometrically sound measure. The present study was designed to fill this gap in the literature so that a valid measure of schadenfreude

can be developed to assess schadenfreude in terms of a superordinate construct comprising three constituents: aggression, rivalry, and justice. The convergent and discriminant validity of the newly developed measures was also assessed. The findings of the current research establish validity evidence for the tripartite motivational model of schadenfreude (Wang et al., 2019).

Hypotheses

We expect that the scores on the *Tripartite Schadenfreude Scale* will correlate positively with the scores on the *Perceived Schadenfreude Scale* (Batool, 2014), which is an indigenously developed unidimensional and psychometrically good measure of perceived schadenfreude. The positive association between these two measures may provide evidence of the convergent validity of the *Tripartite Schadenfreude Scale*.

Affective states of shame and guilt were chosen to discriminate tripartite schadenfreude from these negative emotional states. Marshall (1994) defined shame and guilt as two negative emotions that were considered the same in the sense that a person feels them after doing something wrong, however, these two are different emotions. Shame involves a negative self-appraisal of one's global sense of self (Barrett, 1995; Tangney, 1995), and this negative self-scrutiny arises in response to engaging in some action that is judged to be bad. Shame is experienced when 'bad' behavior is attributed to an internal and unchangeable feature of the person. In this respect, the self is seen as bad, and as such, bad behavior is both inevitable and irremediable. Shame (i.e., 'I am a bad person.'): unfortunately, this results in the person feeling unable to stop the bad behavior that generated the emotion (Marshall, Marshall, Serran & O'Brien, 2009). Guilt, on the other hand, involves the negative appraisal of a specific action (Barrett, 1995; Tangney, 1995). In this case, the person distinguishes between himself/herself as a whole being and the particular action of concern. Essentially persons experiencing guilt say to themselves 'I am not a bad person but I did a bad thing.' Thus, a response of guilt to an unacceptable action, unlike shame, does not involve an appraisal of the person's core identity but rather focuses on stopping problematic behavior (Tangney, 1995). Guilt prompts other-oriented concerns, whereas shame involves self-focus (Marshall et al., 2009). Tangney and Dearing (2002) suggest that events that cause shame and guilt are

social. Thus, feelings of guilt and shame encourage people to act according to socially acceptable and legitimate standards of right and wrong.

Previous literature has provided no evidence of whether schadenfreude experiences lead to feelings of guilt and shame. Guilt and shame originate from moral and self-conscious emotions that obey social norms and traditions. On the other hand, schadenfreude is conceived as a negative and socially unacceptable emotion that is usually kept secret and that can question one's moral state of mind. Schadenfreude is found to be the product of social comparison in rivalry schadenfreude and is exhibited as an effective response to others. Aggression schadenfreude is elicited by a discrepancy between one's group ideal self and the current self by comparing it with the outgroup. Thus, guilt and shame might not occur, as the loss of the outgroup does not outperform the gains of the in-group. Finally, justice schadenfreude is elicited when deservingness requirements are socially and morally acceptable. Thus, justice schadenfreude may not lead to guilt and shame in schadenfroh because guilt and shame occur because of severe violations of moral values, whereas a person who experiences justice schadenfreude considers the misfortune befallen on the other as morally legitimate. Thus, schadenfreude and its components may not be related to shame and guilt and may serve as potentially relevant variables for establishing the discriminant validity of the *Tripartite Schadenfreude Scale* developed in the present study. Based on these arguments, we hypothesize the following:

1. the scores on the *Tripartite Schadenfreude Scale* and its three components will be positively related to scores on the *Perceived Schadenfreude Scale* (Batool, 2014);

2. the *Tripartite Schadenfreude Scale* and its three components will not be related to affective states of shame and guilt.

METHOD

This study comprised of two studies. The first study dealt with the development of TSS in the Urdu language and the establishment of its psychometric properties. In the second study, the factorial structure of the newly developed TSS was reconfirmed through confirmatory factor analysis (CFA). The research was conducted in strict compliance with the ethical guidelines of the American Psychological Association and was monitored by the research ethics committee of the Department of Psychology, University of Sargodha, Pakistan.

Study 1: Development of TSS

Study 1 consisted of two phases. Phase 1 of Study 1 involved the development of an item pool of an indigenous self-report measure of TSS in Urdu. In Phase 2, the final item pool of the TSS was administered to a sample of university students to explore its psychometric properties in terms of reliability, factorial structure, and validity.

– *Phase 1(a): Item pool generation*

To generate an item pool, we followed the guidelines of Burisch (1984). Both qualitative and quantitative approaches have been used to develop the TSS. In the qualitative phase, scientific literature on schadenfreude was reviewed by Wang and colleagues (2019), with an emphasis on developing a deep understanding of the tripartite motivational model of schadenfreude. Aggression, rivalry, and justice schadenfreude items were separately constructed, analyzed, molded, and extracted repeatedly. Phase 1 involved the development of an initial pool of 50 items in several steps. These steps included (i) a review of the pertinent literature, (ii) an in-depth study of the tripartite motivational model of schadenfreude (Wang et al., 2019), and (iii) a review of the existing measures of schadenfreude. The item pool included 17 items on aggression schadenfreude, 17 items on rivalry schadenfreude, and 16 items on justice schadenfreude. A committee of experts reviewed the initial pool of items.

– *Phase 1(b): Review of item pool through committee approach*

In the second part of Phase 1, the committee approach was used to evaluate the item pool of TSS. A committee approach was sought so that each item of the item pool could be evaluated in terms of its language, cultural relevance, appropriateness, comprehension, and understandability of the target population. The committee comprised six faculty members of the Department of Psychology, University of Sargodha; four members had Ph.D. degrees (assistant professors), and two had M.Phil degrees in psychology (lecturers). The committee identified certain items as ambiguous in terms of phrasing, overlapping with other items, and leading or double-barreled items. The amendments of the committee, including phrasal of a couple of items, addition, or obstruction of words, were incorporated into the final items of the scale. Based on the above procedure, a final item pool of 36 items was generated for the assessment of the three sub-forms of schadenfreude. These sub-forms include aggression

schadenfreude with 13 items (item no.1-13), rivalry schadenfreude with 13 items (item no.14-26), and justice schadenfreude with 10 items (item no.27-36).

– *Phase 2(a): Content validity*

Views from experts proficient in the psychometrics field were obtained to assess the content validity of the TSS. The major objective of obtaining opinions from experts was to ensure content validity. The experts recommended that the items be included in the final scale, which was in line with the theoretical conception of the tripartite motivational model.

Participants and procedure: to establish the content validity of the TSS, four faculty members of the Department of Psychology, University of Sargodha, were given an article by Wang and colleagues (2019), who explained the tripartite motivational model of schadenfreude in detail. They were requested to go through the article so that they could have a better understanding of the theory behind the development of the TSS and could readily understand rivalry, justice, and aggression schadenfreude. They were then asked to rate each item for its correspondence with one of the three facets/sub-forms of schadenfreude (rivalry, aggression, and justice) on a 4-point (1-4) rating scale (1 = not relevant, 2 = somewhat relevant, 3 = quite relevant, and 4 = highly relevant).

The content validity indices for items (I-CVIs) and scale (S-CVI) were calculated. Items with I-CVIs lower than .78 were excluded (Lynn, 1986). In this way, six items from the TSS were omitted, and 30 items were retained for the final scale, which were used for the evaluation of additional psychometric properties. Finally, TSS was converted into a self-report measure consisting of a 5-point rating scale (1-5), where 1 = not at all and 5 = always. The committee also consensually finalized the items falling in each facet/sub-form of schadenfreude.

– *Phase 2(b): Assessment of psychometric properties of the TSS*

Phase 2 involved assessment of the psychometric properties of the TSS in terms of its factorial structure, reliability, and construct validity.

Study 1: Sample

To determine the appropriate sample size, we conducted a power analysis using the tool developed by Preacher and Coffman (2006). The findings indicated that with a power of .90 and $\alpha = .01$, the null hypothesis that the Root Mean Square Error of Approximation (RMSEA) = .00, against the alternate hypothesis that RMSEA = .05, with $df = 393$, could be reliably tested with a sample size of 116. To be more cautious, we recruited a purposive sample of Study 1 comprising ($N = 300$) undergraduate students from a large public sector Pakistani university. The sample had an almost equal representation of both sexes. The age of the sample ranged from 18 to 25 years ($M = 21.68$ years; $SD = 2.78$ years). The inclusion criteria dictated that the sample of the study was restricted to full-time university students with an age range of 18-25 years enrolled in BS and master's programs. Postgraduate students, part-time students, and students doing jobs were not included in the study sample.

Study 1: Instruments

All measures used in Study 1 were self-reported psychometrically sound instruments in Urdu; the details are as follows.

– *TSS Initial Item Pool.* The final item pool of the TSS consisting of 30 items on a 5-point Likert rating scale ranging from 1 = strongly agree to 5 = strongly disagree was administered to the participants. There were no negative items. Scores were obtained by calculating the sum of the scores for each item on the scale. There were three subscales: aggression schadenfreude (item no. 1-12), rivalry schadenfreude (item no. 13-24), and justice schadenfreude (item no. 25-30). Possible scores ranged from 30 to 150, with higher scores indicating greater schadenfreude.

– *Perceived Schadenfreude Scale (PSS).* The PSS (Batool, 2014) was used to measure perceived schadenfreude. It comprises 28 items with a 5-point Likert-type scale ranging from 1 = strongly disagree to 5 = strongly agree. The scale showed a satisfactory alpha coefficient of .82 (Batool 2014). Possible scores ranged from 28 to 140, with high scores indicating greater schadenfreude.

Study 1: Procedure

After finalization of the scales and permission from the authors of all the scales, informed consent and demographic forms were constructed, and the final draft of

the questionnaire form was composed. Participants' written informed consent was also obtained, and it was assured that the information would be used only for research purposes and that full confidentiality would be maintained. The participants were provided with a questionnaire booklet in a paper-pencil format and were briefed that their responses would be useful for scientific knowledge; however, they had the right to withdraw from participation in the study. They were asked to provide honest responses. The response rate was 82.64%.

Study 2: Replication of CFA and validation of TSS

In Study 2, the TSS (finalized as a result of Study 1) and the *State of Shame and Guilt Scale* (*SSGS*; Zia, 2018) were administered to a purposive sample of (*N* = 219) undergraduate students at the University of Sargodha. In this study, the factorial structure of the TSS was replicated using confirmatory factor analysis (CFA). CFA was undertaken instead of EFA because the purpose of factor analysis was to confirm the factor structure of the TSS items pool in accordance with the tripartite motivational model of schadenfreude (Wang et al., 2019) instead of exploring the factor structure. The items in the item pool of the TSS were developed in accordance with the three forms of schadenfreude as per the tripartite motivational model of schadenfreude, and were content validated through an empirical procedure. This study fulfilled the requirements i.e the assumptions and critical steps of conducting CFA as it is rooted in hypothesized measurement model based on theory and prior research. While describing the comparison between EFA and CFA, it has been argued by Brown (2006) that CFA is often used in scale development and validation when there are established theoretical and empirical grounds for construct being measured. Orcan (2018) stated that CFA can be initiated as factor analysis technique in scale construction studies if there is hypothetically tested and known relationship between factors. According to Maltby (2002) CFA is more powerful and suitable than EFA for a theory-based construct as it provides the researcher a tool to reject or accept a theory. An examination of relevant literature reveals several studies that utilized CFA for scale development grounded in theoretical frameworks. For instance, Schaufeli and colleagues (Schaufeli, Bakker & Salanova, 2006) employed CFA in the development of the *Short Utrecht Work Engagement Scale* (*UWES*-9) in their

cross-national research spanning ten countries. Therefore, we ran a CFA to test whether the content validated items loaded on their corresponding factors as per the hypothesized measurement model. In addition to confirming the factorial structure of TSS, evidence for the discriminant validity of TSS was also established.

Study 2: Sample

To determine the appropriate sample size for Study 2, we conducted a power analysis using a tool developed by Preacher and Coffman (2006). The results suggest that the null hypothesis that the Root Mean Square Error of Approximation (RMSEA) = .00, against the alternate hypothesis that RMSEA = .05, with *df* = 401, could reliably be tested with a sample size of 114 with a power of 90% and α = .01. More cautiously, we recruited a purposive sample of Study 2 comprising (*N* = 219) undergraduate students (111 girls and 108 boys) from a large public sector Pakistani university during the spring semester of 2020. Participants' ages ranged from 18-25 (*M* = 22.48 years, *SD* = 2.89 years). The inclusion criteria dictated that the sample of the study was restricted to full-time students of the University of Sargodha with an age range of 18-25 years enrolled in BS and master's programs. Postgraduate students, part-time students, and students doing jobs were not included in the study sample.

Study 2: Instruments

All measures used in Study 2 were self-reported psychometrically sound instruments in Urdu; the details are as follows.

– *Tripartite Schadenfreude Scale*. The TSS, developed as a result of the CFA in Study 1, was used to measure schadenfreude. The scale consisted of 30 items. The three sub-forms are as follows: aggression schadenfreude contains 12 items (item no. 1-12), rivalry schadenfreude has 12 items (item no. 13-24), and justice schadenfreude has six items (item no. 25-30). The scale showed a good alpha reliability. The reliability of TSS was also satisfactory with an alpha coefficient of .97, while its subscales were aggression schadenfreude (α = .86), rivalry schadenfreude (α = .87), and justice schadenfreude (α = .73).

– *State Shame and Guilt Scale (SSGS)*. The Urdu-translated

version (Zia, 2018) of the *State Shame and Guilt Scale* (*SSGS*; Cavalera et al., 2017) was used to measure shame and guilt. The scale consisted of 10 items. It is rated on a 5-point Likert scale. A high score reflected a high degree of guilt or shame. The Urdu version of the SSGS has shown good reliability, with an alpha coefficient of .76 (Zia, 2018).

Study 2: Procedure

All the scales required for the validation of the indigenously developed TSS were distributed among the study participants. First, written consent was obtained from the participants, and instructions were given regarding filling out the scales. Participants were assured that their information would be used only for the purpose of the study and kept confidential.

RESULTS

Study 1 was conducted in two phases. In Phase 1, the content validity of the TSS was ensured. In Phase 2, confirmatory factor analysis, reliability, and correlations were computed to ensure the validity of the TSS.

Study 1: Phase 1 - Content validity

According to Lynn's (1986) guidelines, items with a content validity index lower than .78 were excluded. In this way, six items from the item pool of the TSS were excluded, and 30 items were retained for the final scale and used for additional psychometric properties. Moreover, the averaging approach (S-CVI/Ave) was used to calculate the S-CVI of TSS. Waltz and colleagues (Waltz, Strickland & Lenz, 2005) and Lynn (1986) recommended that the S-CVI/Ave should be .90 or higher. In the current study, S-CVI/Ave was .95, which indicated good content validity of the TSS.

Study 1: Phase 2 - Confirmatory factor analysis of TSS

The TSS was subjected to CFA using maximum likelihood estimation to discern whether the same three-factor structure

could be replicated, as found in Study 1. An assessment of the multivariate normality of TSS items revealed that the data met the assumptions required for maximum likelihood (ML) estimation in confirmatory factor analysis (CFA). Skewness and kurtosis statistics for each item fell within acceptable ranges, indicating univariate normality. Mardia's multivariate kurtosis test confirmed that the data did not significantly deviate from multivariate normality. Additionally, Q-Q plots showed that the items closely followed the expected normal distribution, and Mahalanobis distance calculations did not identify any significant multivariate outliers. These results confirm that the items of TSS were suitable for ML estimation in CFA. The CFA was computed through the AMOS-24 to confirm whether the item pool of the TSS yields a three-factor structure in consonance with the tripartite motivational model of schadenfreude. The fit indices of the competing factorial model of the scale are presented in Table 2, and the standardized loadings of the items are presented in Table 3.

Table 2 shows the stepwise model fit indices for the CFA of the TSS. The first measurement model involved a single-factor model and demonstrated a poor fit with the data. The second model was a two-factor second-order model with schadenfreude as the second-order factor, whereas justice and rivalry plus aggression were first-order factors. This model demonstrated a significantly better fit to the data than Model 1 did. The third model is the proposed measurement model, which specifies a three-factor second-order model with TSS as the second-order factor and aggression schadenfreude (12 indicators), rivalry schadenfreude (12 indicators), and justice schadenfreude (six indicators) as first-order factors with independent error variances. This model again demonstrated a superior fit to the data compared to Model 2. An inspection of the model fit indices suggested further room for improvement of the model fit. Therefore, we allowed a few error variances in the same latent factor to covary according to the modification indices. The fit indices of this model are reported in Model 4 in Table 2, which suggests that Model 4 demonstrated a very good fit to the data, and it was superior to Model 3 as the chi-square difference test between Model 3 and Model 4 was significant. Model 4 depicts the results of the confirmatory factor analysis, where 30 indicators loaded on their respective first-order factors, and the three first-order factors converged on the superordinate construct of the TSS.

Table 3 shows the standard factor loadings of the second-order confirmatory factor analysis of TSS. In this study, a three-factor structure was obtained using CFA. The first

Table 1 – Expert ratings on a 36-item pool of the TSS (Study 1)

Item no.	Expert 1	Expert 2	Expert 3	Expert 4	Agreements	Item CVI
1	X	X	X	X	4	1.0
2	X	X	X	X	4	1.0
3	X	X	X	X	4	1.0
4	X	X	X	—	3	.75
5	X	X	X	X	4	1.0
6	X	X	X	—	3	.75
7	X	X	X	X	4	1.0
8	X	X	X	X	4	1.0
9	X	X	X	X	4	1.0
10	X	X	X	X	4	1.0
11	X	X	X	X	4	1.0
12	X	X	X	X	4	1.0
13	X	X	X	X	4	1.0
14	X	X	X	X	4	1.0
15	X	X	X	X	4	1.0
16	X	X	X	X	4	1.0
17	X	X	X	X	4	1.0
18	X	X	X	X	4	1.0
19	X	X	X	X	4	1.0
20	X	—	X	X	3	.75
21	X	X	X	X	4	1.0
22	X	X	X	X	4	1.0
23	—	X	X	X	3	.75
24	X	X	X	X	4	1.0
25	X	—	X	X	3	.75
26	X	X	X	X	4	1.0
27	X	X	X	X	4	1.0
28	X	X	X	X	4	1.0
29	X	X	—	X	3	.75

Item no.	Expert 1	Expert 2	Expert 3	Expert 4	Agreements	Item CVI
30	X	X	X	—	3	.75
31	X	X	X	X	4	1.0
32	X	X	X	X	4	1.0
33	X	X	X	X	4	1.0
34	X	X	X	X	4	1.0
35	X	X	X	X	4	1.0
36	X	X	X	X	4	1.0
Proportion relevant					*Mean* I-CVI = .95	
	.95	.95	.95	.95	S-CVI/UA = .83	
					Mean Expert Opinion = .90	

Legenda. I-CVI = item-level content validity index; S-CVI/UA = scale-level content validity index/universal aagreement calculation method.

Table 2 – Stepwise model fit for CFA of item pool of TSS (Study 1, N = 300)

Models	χ^2	df	Fit indices				$\Delta\chi^2$	Δdf
			CFI	SRMR	RMSEA	p_{χ^2}		
Model 1 (30 items, single-factor)								
	1300.24	404	.78	.07	.09	<.001	–	–
Model 2 (30 items, two-factor)								
	950	403	.86	.06	.07	<.001	350.24***	1
Model 3 (30 items, three-factor, independent error variances)								
	844.22	402	.88	.05	.06	<.01	105.78***	1
Model 4 (30 items, three-factor, error variances allowed to covary)								
	734.24	393	.95	.048	.05	<.001	109.98***	9

Legenda. CFI = Comparative Fit Index; SRMR = Standardized Root Mean Square Residual; RMSEA = Root Mean Square Error of Approximation; *df* = degree of freedom.
*** $p < .001$

Table 3 – Summary of confirmatory factor analysis of item pool of TSS (Study 1, N = 300)

Items	Rivalry	Aggression	Justice
	.98 (.93)		
15	.68		
16	.64		
17	.72		
18	.61		
19	.64		
20	.64		
21	.60		
22	.63		
24	.69		
26	.65		
27	.65		
28	.61		
		.92 (.94)	
1		.63	
2		.60	
3		.61	
5		.62	
7		.67	
8		.68	
9		.64	
10		.64	
11		.61	
12		.65	
13		.66	
14		.68	

continued

Items	Rivalry	Aggression	Justice
			.67 (.73)
31			.71
32			.80
33			.73
34			.74
35			.67

Note. Standardized factor loadings of first-order loading are bold. The alpha coefficients of each factor are shown in parenthesis.

factor measures rivalry and contains 12 items; the second factor measures aggression, and it also comprises 12 items; and the third factor measures justice and comprises six items.

Figure 1 shows the factor structure of the three sub-forms of the TSS: aggression, rivalry, and justice schadenfreude. The item loadings for all first-order factors remained at ≥.60, whereas the second-order factor loadings remained at ≥.68. TSS explained 85% of the variance in aggression, 97% in rivalry, and 44% in justice. To improve the model fit, a few error variances on the same first-order factor were allowed to covary, as suggested by the modification indices.

Table 4 shows the psychometric properties of the study constructs. The alpha coefficients for all scales ranged from .70 to .93, which indicated satisfactory internal consistency. Table 4 also shows that all subscales were positively correlated with each other.

Study 2: Validation of the factorial structure of the TSS

In Study 2, the same procedure was applied to assess the multivariate normality of TSS items, confirming that the data met the assumptions required for maximum likelihood (ML) estimation in confirmatory factor analysis (CFA). The findings of the CFA demonstrated that the data of Study 2 fit well with the hypothesized measurement model of the TSS, which validated the factorial structure of the TSS established in Study 1. The model fit indices indicated a good fit (χ^2 = 869.70, df = 401, p<.001; CFI = .96; SRMR = .041; RMSEA = .048; p = .07). The results of confirmatory factor analysis revealed that the 30 indicators loaded on their respective first-order factors, and the three first-order factors converged on the superordinate construct of schadenfreude.

Table 5 shows the standard factor loadings of the second-order confirmatory factor analysis of the TSS. All indicators had a standardized factor loading ≥.40, which revealed that items of various factors of schadenfreude had unique contributions to the operationalization of this construct. In this study, a three-factor structure was obtained using CFA. The first factor measures rivalry and contains 12 items; the second factor measures aggression, and it also comprises 12 items; and the third factor measures justice and comprises six items.

Figure 2 shows the factor structure of the three subscales of the TSS: aggression, rivalry, and justice schadenfreude. Item loadings for all first-order factors remained ≥.40, whereas second-order factor loadings remained ≥.90. TSS explained 94% variance in aggression, 96% in rivalry, and 85 % in justice. To improve the model fit, a couple of error variances on the same first-order factor were allowed to covary, as suggested by the modification indices. First, the error terms of items 4 and 5 of the rivalry schadenfreude were allowed to covary because both items shared a common theme

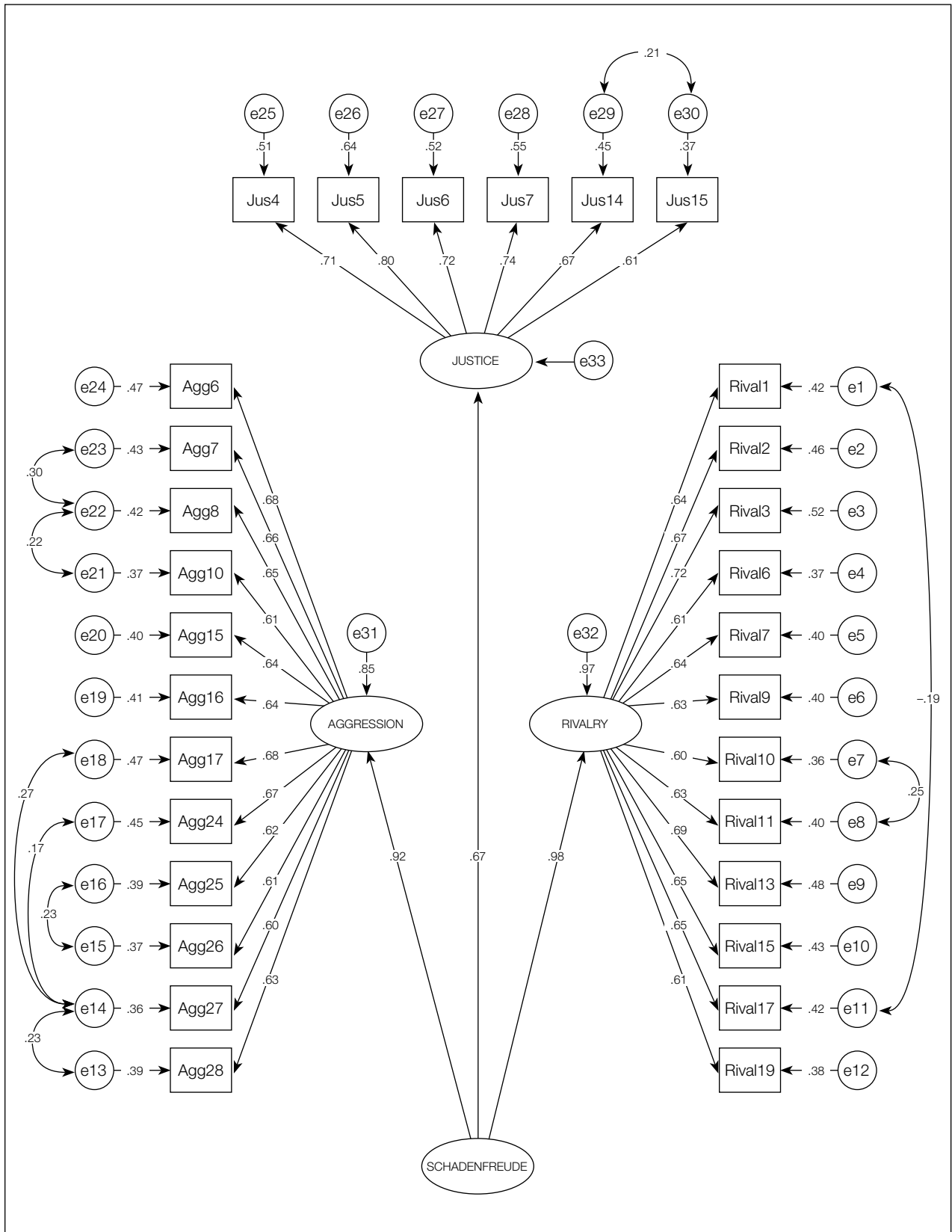**Figure 1** – Factor structure of the three sub-forms of the TSS

Table 4 – Means, standard deviations, alpha reliabilities and correlations among the variables of Study 1 (N = 300)

Variables	1	2	3	4	5	M	SD	α	Sk[a]	Ku[b]
1. PSS	—	.60**	.69**	.69**	.74**	59.65	12.02	.93	.09	−.15
2. ASS		—	.81**	.54**	.93***	24.81	10.1	.86	.77	.30
3. RSS			—	.59**	.94***	23.89	9.60	.87	.68	−.27
4. JSS				—	.74**	13.61	5.7	.73	.48	.14
5. TSS					—	63.32	22.7	.97	.63	−.10

Legenda. PSS = *Perceived Schadenfreude Scale*; ASS = *Aggression Schadenfreude Scale*; RSS = *Rivalry Schadenfreude Scale*; JSS = *Justice Schadenfreude Scale*; TSS = *Tripartite Schadenfreude Scale*.

Note. [a]Standard error of skewness = .13; [b]standard error of kurtosis = .27.
** $p<.01$; *** $p<.001$

Table 5 – Summary of confirmatory factor analysis of TSS (Study 2, N = 219)

Items	Rivalry	Aggression	Justice
	.98 (.93)		
Riv 1	.49		
Riv 2	.78		
Riv 3	.77		
Riv 4	.54		
Riv 5	.76		
Riv 6	.70		
Riv 7	.47		
Riv 8	.44		
Riv 9	.72		
Riv 10	.40		
Riv 11	.61		
Riv 12	.61		

continued

Items	Rivalry	Aggression	Justice
		.97 (.94)	
Agg 1		.43	
Agg 2		.65	
Agg 3		.69	
Agg 4		.64	
Agg 5		.58	
Agg 6		.58	
Agg 7		.62	
Agg 8		.77	
Agg 9		.66	
Agg 10		.61	
Agg 11		.77	
Agg 12		.55	
			.92 (.73)
Jus 1			.61
Jus 2			.50
Jus 3			.50
Jus 4			.53
Jus 5			.82
Jus 6			.54

Legenda. Riv = rivalry schadenfreude; Agg = aggression schadenfreude; Jus = justice schadenfreude.

Note. Standardized factor loadings of first-order loading are bold. The alpha coefficients of each factor are shown in parenthesis.

Figure 2 – Factor structure of the three subscale of the TSS
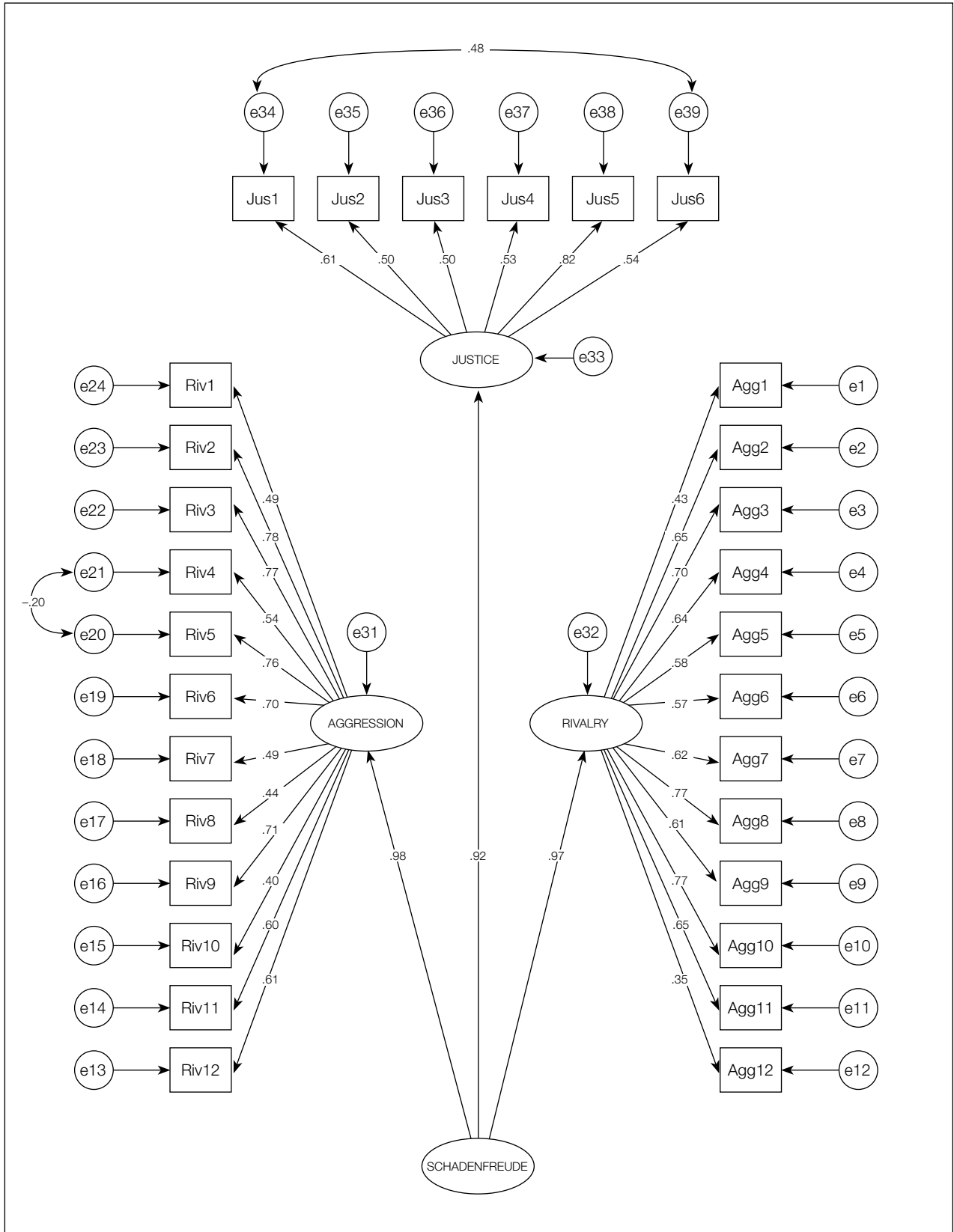
related to dominance and the desire to establish superiority over rivals. They capture different aspects of dominance and satisfaction derived from exerting power over rivals who attempt to dominate or challenge one's position. Second, the error terms of items 2 and 6 of justice schadenfreude were allowed to covary because both items reflect different aspects of schadenfreude but share a common theme related to the satisfaction or pleasure derived from witnessing negative outcomes for individuals who have engaged in deceptive or wrongful behavior.

Table 6 shows the psychometric properties of the study constructs. The alpha coefficients for all scales ranged from .70 to .93, which indicated satisfactory internal consistency. Table 6 also depicts the non-significant relationships between states of guilt and shame with the three sub-forms of TSS.

DISCUSSION

This study aimed to develop a comprehensive and indigenous measure of schadenfreude based on a tripartite motivational model. The literature provides a basis for this model and a theoretical foundation for the item generation of the scale. The extensive study and analyses of this model have paved the way for item generation. The cultural and contextual conditions of the local subjects, their characteristics, and their behavioral patterns were also considered while developing the item pool of the scale. The schadenfreude construct was measured in Urdu, the language of the targeted population. To the best of our knowledge, no psychometrically sound measure of schadenfreude in consonance with the tripartite motivational model has been developed.

The *Tripartite Schadenfreude Scale* (*TSS*) underwent the standard content validity procedures recommended by Waltz et al. (2005) and Lynn (1986). Thirty out of 36 items with I-CVIs greater than .80 were retained, resulting in an excellent scale-level content validity index (S-CVI). TSS is considered an excellent measure for assessing tripartite schadenfreude because of its strong content validity.

Confirmatory factor analysis confirmed the three-factor structure of the *Tripartite Schadenfreude Scale* (*TSS*), aligned with the tripartite model of schadenfreude. The final model consisted of 30 items, with high internal consistency and reliability. The factor structure was successfully replicated in an independent sample, demonstrating a good fit for the proposed TSS measurement model.

Table 6 – Means, standard deviations, alpha reliabilities and correlations among the variables of Study 2 (N = 219)

Variables	1	2	3	4	5	M	SD	α	Sk[a]	Ku[b]
1. TSS	—	.83**	.86**	.09	.12	22.63	8.90	.86	.64	−.59
2. RSS			. 66**	.08	.11	22.53	9.40	.87	.87	−.91
3. JSS				.07	.07	13.33	7.75	.75	.33	−.42
4. SGS					.75**	15.04	4.65	.70	−.06	−.74
5. SSS					—	14.92	5.69	.83	−.17	−.90

Legenda. TSS = *Tripartite Schadenfreude Scale*; RSS = *Rivalry Schadenfreude Scale*; JSS = *Justice Schadenfreude Scale*; SGS = *State of Guilt Scale;* SSS = *State of Shame Scale*.

Note. [a]Standard error of skewness = .12; [b]standard error of kurtosis = .26.

** *p*<.01; *** *p*<.001

The first-factor, aggression schadenfreude, measures the level of experience of schadenfreude due to a threat to in-group identity from the outgroup. Aggression schadenfreude arises from the sense of social identity and group membership. This occurs when individuals derive pleasure from witnessing the misfortune or suffering of out-group members and those who do not belong to their identified group. It is driven by a hostile or competitive mindset towards the out-group and can involve feelings of superiority, dominance, or satisfaction over their downfall. Thus, it can be defined as "on a superior group's failure, the spontaneous pleasure felt by the in-group members to enhance in-group identity is known as aggression schadenfreude".

Items on the second factor represent rivalry schadenfreude. Rivalry schadenfreude primarily focuses on one's position in social comparison. It occurs when individuals experience pleasure from the misfortunes or failures of others who are seen as rivals or competitors. The primary motivation behind rivalry schadenfreude is to enhance one's self-esteem or self-worth by highlighting superiority or relative success in comparison to the rival. It can be defined as "the pleasure felt by exploiting others whenever possible in social comparison". It stems from the preference for advantageous inequity, which means that inequity is preferred when you are benefitting and not when others are on benefited end. It arises from an earlier sense of self-identity.

The items on the third factor measure justice schadenfreude, which refers to the social phenomenon of deservingness. Justice schadenfreude is other-oriented and revolves around concerns about social justice and fairness. It occurs when individuals derive satisfaction from seeing individuals who violate social norms or principles of fairness receive appropriate punishment or consequences for their actions. The primary goal of justice schadenfreude is to ensure that justice is served and that those who transgress societal rules face retribution. The person who experiences justice schadenfreude wants to confirm whether the individual who violates social norms, has received punishment and if the violator received punishment, what is his/her affective state. Thus, the affective state of the violator is a source of pleasure for schadenfroh (Jensen, 2012).

In summary, aggression schadenfreude is driven by intergroup dynamics, rivalry schadenfreude focuses on self-enhancement through social comparison, and justice schadenfreude is motivated by a desire for fairness and punishment of social violators. These sub-forms highlight the diverse motivations and contexts in which schadenfreude can manifest. Accordingly, in the light of the factor structure of the TSS, tripartite schadenfreude can be defined as "the feeling of pleasure on the misfortune of others, while appraising another's misfortune in terms of personal gains in enhancing one's social identity, self-identity, and belief in social justice". Intergroup aggression emerges due to social identity; self-evaluation emerges due to social comparison, and the sense of deservingness emerges due to belief in a just world. This definition seems to be comprehensive as it involves all three factors, which operationalize tripartite schadenfreude.

To establish the evidence for the convergent validity of the TSS, the scores on the TSS were correlated with scores on PSS. The results show that the TSS has a strong positive correlation with the PSS. Correlation analysis depicts that aggression, rivalry, and justice schadenfreude sub-scales are significantly and strongly positively related to the PSS. Perceived schadenfreude involves the tendency to experience schadenfreude following perceiving another's misfortune (Naila, 2014). This confirms the conceptual relevance of both measures. The findings showed that the nomological network of schadenfreude measured by both scales is also comparable.

The discriminant validity of the TSS was established against the state of shame and guilt as we reasoned that people who are high on schadenfreude may not experience shame or guilt on the misfortune of others; rather they may feel pleasure in it. As hypothesized, the results have shown that TSS and its components have a non-significant correlation with the state of shame and the state of guilt. There are several reasons to support the absence of an association between schadenfreude and feelings of shame and guilt. Firstly, schadenfreude is primarily directed toward others' misfortune or suffering. It involves deriving pleasure or satisfaction from witnessing the downfall or failures of others. In contrast, shame and guilt are predominantly self-focused emotions that arise from one's own perceived shortcomings, mistakes, or transgressions. The focus of schadenfreude on external events and others' experiences makes it less likely to be closely related to the internal states of shame and guilt.

Secondly, schadenfreude is often characterized by a lack of self-reflection or introspection. Individuals experiencing schadenfreude may not critically evaluate their behavior or actions concerning others' suffering. In contrast, shame and guilt involve self-evaluation and an awareness of personal responsibility for the negative outcomes or harm caused

to others. The absence of self-reflection in schadenfreude further reduces the likelihood of a strong correlation with shame and guilt.

Thirdly, schadenfreude is typically associated with positive affect or pleasure derived from others' misfortune. Shame and guilt, on the other hand, are negative emotions that arise from a sense of wrongdoing or moral transgressions. These opposing emotional valences make it less likely for schadenfreude to be strongly correlated with shame and guilt.

While there may be some cases where schadenfreude and feelings of shame or guilt co-occur, such as when individuals feel shame or guilt about experiencing schadenfreude itself, the general expectation is that schadenfreude is not strongly correlated with shame and guilt due to its different focus, lack of self-reflection, and opposing emotional valence.

CONCLUSION

Schadenfreude is an important concept in social psychology. The present study developed a psychometrically sound measure of schadenfreude based on the tripartite motivational model of schadenfreude. Tripartite schadenfreude involves appraising another's misfortune as a means to achieve personal goals related to social identity, self-identity, deservingness, and feeling pleasure. The TSS developed in the present study has demonstrated that it was content valid as it sufficiently covered various aspects of the conceptual definition of schadenfreude. Furthermore, evidence for the construct validity of TSS has also been established through factor analysis and convergent and discriminant validity. TSS and its various subscales also demonstrated satisfactory levels of internal consistency.

Limitations and suggestions

The present study has limited generalizability as the data were collected from some Departments of the University of Sargodha, so to enhance the external validity, further research should be conducted on large and diverse samples. As all the scales were self-reported measures, therefore mono-method bias and social desirability can be a potential threat to internal validity. Schadenfreude is generally viewed as a negative emotion. People did not admit that they have felt such emotions. Therefore, social desirability may also have

interrupted the reporting of socially undesirable behaviors. It might be better to use vignettes instead of simple items to reduce social desirability. However, vignettes take a longer time for participants' responses and there is a risk of fatigue. Moreover, vignettes may not have sound psychometric properties. The self-report measure of TSS has been developed and validated in the Urdu language. Therefore, to apply it across regions and cultures, it must be translated into other languages as well. It will also help establish more evidence for the psychometric soundness of the TSS.

The present study provides a novel self-reporting measure of schadenfreude in the field of psychology, established its nomological network, and explored some of its correlates. It is recommended to researchers to further elaborate on the nomological structure of schadenfreude by applying it to samples of different regions. It is highly encouraged to translate TSS into other languages so that it can be used in diverse cultures and regions. Moreover, research studies while operationalizing TSS on ethnic minorities, with different variables should be carried out to enhance generalizability. Experimental research should be carried out to infer causal relationships of schadenfreude with social constructs. Furthermore, TSS should be cross-culturally validated in future studies.

Implications of the present study

Tripartite schadenfreude can improve people's view related to schadenfreude and their acceptance of it as a normal phenomenon because tripartite schadenfreude helps people to enhance their social identity, self-identify, and belief in social justice. This life improvement also leads to better flourishing and overall satisfaction with one's life. The development of the TSS as a psychometrically sound measure of schadenfreude may open new avenues of research on this important personal strength. In future studies, this scale may be used for the assessment of schadenfreude and its three aspects with various correlates, which will help expand the nomological network of schadenfreude. Furthermore, cross-cultural research on schadenfreude may reflect whether it is a universal personal strength or if it may have different factorial structures across different cultures. Thus, the development of this scale may contribute to the fields of positive and cross-cultural psychology.

Schadenfreude has been a notoriously difficult concept to

measure, and the present study would support that conclusion. As has been stated before, it may be that schadenfreude defies the logic of being a singular unit, but is best understood by its three aspects. The TSS as developed in the present study has shown that it adequately covers the conceptual definition of tripartite schadenfreude and has shown construct validity. The present TSS provides a workable instrument to continue the advancement of our knowledge of this important construct.

References

BAREN, A. (2017). *Why some take pleasure in other people's pain: The role of attachment, competition, and cooperation on schadenfreude.* Unpublished doctoral dissertation, City University of New York.

BARRETT, K.C. (1995). A functionalist approach to shame and guilt. In J.P. Tangney & K.W. Fischer (Eds.), *Self-conscious emotions: The psychology of shame, guilt, embarrassment, and pride.* New York, NY: Guilford Press.

BATOOL, N. (2014). *Schadenfreude among adolescents: Role of social comparison, self-esteem, episodic envoy and belief in just world.* Unpublished MPhil thesis, National Institute of Psychology, Quaid-i-Azam University, Islamabad, Pakistan.

BROWN, T.A. (2006). *Confirmatory factor analysis for applied research.* New York, NY: Guilford.

CAVALERA, C., PEPE, A., ZURLONI, V., DIANA, B., REALDON, O., & JIANG, R. (2017). A short version of the State Shame and Guilt Scale (SSGS-8). *TPM-Testing, Psychometrics, Methodology in Applied Psychology*, 24 (1), 99-106.

CIKARA, M., BOTVINICK, M.M., & FISKE, S.T. (2011). Us versus them: Social identity shapes neural responses to intergroup competition and harm. *Psychological Science*, 22 (3), 306-313. doi.org/10.1177/0956797610397667

DEMIR, E., & RIGONI, U. (2017). You lose, I feel better: Rivalry between soccer teams and the impact of schadenfreude on stock market. *Journal of Sports Economics*, 18 (1), 58-76. doi.org/10.1177/1527002514551801

GREGORY, S. (2014). *Psychological testing: History, principles, and applications.* London, England: Pearson

HAMLIN, J.K., MAHAJAN, N., LIBERMAN, Z., & WYNN, K. (2013). Not like me is bad: Infants prefer those who harm dissimilar others. *Psychological Science*, 24 (4), 589-594. doi.org/10.1177/0956797612457785

HAMLIN, J.K., & WYNN, K. (2011). Young infants prefer prosocial to antisocial others. *Cognitive Development*, 26 (1), 30-39. doi.org/10.1016/j.cogdev.2010.09.001

HARELI, S., & WEINER, B. (2002). Dislike and envy as antecedents of pleasure at another's misfortune. *Motivation and Emotion*, 26 (4), 257-277. doi.org/10.1023/A:1022818803399

HEIDER, F. (1958). *The psychology of interpersonal relations.* Wiley.

JENSEN, K. (2012). Who cares? Other-regarding concerns – decisions with feeling. In P. Hammerstein, & J.R. Stevens (Eds.), *Evolution and the mechanisms of decision making.* Cambridge, MA: MIT Press.

LEACH, C.W., SPEARS, R., BRANSCOMBE, N.R., & DOOSJE, B. (2003). Malicious pleasure: Schadenfreude at the suffering of another group. *Journal of Personality and Social Psychology*, 84 (5), 932-943. psycnet.apa.org/doi/10.1037/00223514.84.5.932

LERNER, M.J., & MILLER, D.T. (1978). Just world research and the attribution process: Looking back and ahead. *Psychological Bulletin*, 85 (5), 1030-1041. doi.org/10.1037/0033-2909.85.5.1030

LYNN, M.R. (1986). Determination and quantification of content validity. *Nursing Research*, *35*, 382-385.

MALTBY, J. (2002). The Age Universal I-E Scale-12 and orientation toward religion: Confirmatory factor analysis. *The Journal of Psychology: Interdisciplinary and Applied*, *136* (5), 555-560. dx.doi.org/10.1080/00223980209605550

MARSHALL, W.L. (1994). Treatment effects on denial and minimization in incarcerated sex offenders. *Behavior Research and Therapy*, *32*, 559-564.

MARSHALL, W.L., MARSHALL, L.E. SERRAN, G.A., & O'BRIEN, M.D. (2009). Self-esteem, shame, cognitive distortions and empathy in sexual offenders: Their integration and treatment implications. *Psychology, Crime* & *Law*, *15* (2), 217-234.

NOBES, G., PANAGIOTAKI, G., & PAWSON, C. (2009). The influence of negligence, intention, and outcome on children's moral judgments. *Journal of Experimental Child Psychology*, *104* (4), 382-397. doi.org/10.1016/j.jecp.2009.08.001

ORCAN, F. (2018). Exploratory and confirmatory factor analysis: Which one to use first? *Journal of Measurement and Evaluation in Education and Psychology*, *9* (4), 414-421. doi.org/10.2103//epod.394323

OUWERKERK, J.W., & VAN DIJK, W.W. (2008, June 10-14). *"Don't mention the war": Social identification and salience of rivalry as determinants of intergroup schadenfreude*. Paper presented at the 15th General Meeting of the European Association for Experimental Social Psychology. Opatija: Croatia.

PIETRASZKIEWICZ, A. (2013). Schadenfreude and just world belief. *Australian Journal of Psychology*, *65* (3), 188-194. doi.org/10.1111/ajpy.12020

PISKORZ, J.E., & PISKORZ, Z. (2009). Situational determinants of envy and schadenfreude. *Polish Psychological Bulletin*, *40* (3), 137-144. psycnet.apa.org/doi/10.2478/s10059-009-0030-2

PREACHER, K.J., & COFFMAN, D.L. (2006). *Computing power and minimum sample size for RMSEA [Computer software]*. Retrieved from: http://quantpsy.org/

SAGHER, S. (2017). *Correlates of internet addiction and sensation-seeking behavior among adolescents and young adults*. Unpublished master's thesis, Department of Psychology, University of Sargodha, Pakistan.

SCHAUFELI, W.B., BAKKER, A.B., & SALANOVA, M. (2006). The measurement of work engagement with a short questionnaire: A cross-national study. *Educational and Psychological Measurement*, *66* (4), 701-716.

SMITH, R.H., POWELL, C.A., COMBS, D.J., & SCHURTZ, D.R. (2009). Exploring the when and why of schadenfreude. *Social and Personality Psychology Compass*, *3* (4), 530-546. doi.org/10.1111/j.1751-9004.2009.00181.x

SMITH, R.H., & VAN DIJK, W.W. (2018). Schadenfreude and Gluckschmerz. *Emotion Review*, *10* (4), 293-304. doi.org/10.1177/1754073918765657

STEINBEIS, N., & SINGER, T. (2013). The effects of social comparison on social emotions and behavior during childhood: The ontogeny of envy and schadenfreude predicts developmental changes inequity-related decisions. *Journal of Experimental Child Psychology*, *115* (1), 198-209. doi.org/10.1016/j.jecp.2012.11.009

TAJFEL, H., & TURNER, J.E. (1986). The social identity theory of intergroup behavior. In S. Worchel & W. Austin (Eds.), *Psychology of intergroup relations*. Nelson-Hall.

TANGNEY, J.P. (1995). Shame and guilt in interpersonal relationships. In J.P. Tangney, & K.W. Fischer (Eds.), *Self-conscious emotions: The psychology of shame and guilt, embarrassment, and pride*. New York, NY: Guilford.

TANGNEY, J.P., & DEARING, R.L. (2002). *Shame and guilt*. New York, NY: Guilford.

VAN DE VEN, N., HOOGLAND, C.E., SMITH, R.H., VAN DIJK, W.W., BREUGELMANS, S.M., & ZEELENBERG, M. (2015). When envy leads to schadenfreude. *Cognition and Emotion*, *29* (6), 1007-1025.

VAN DIJK, W.W., GOSLINGA, S., & OUWERKERK, J.W. (2008). Impact of responsibility for misfortune on schadenfreude and sympathy: Further evidence. *The Journal of Social Psychology*, *148* (5), 631-636. doi.org/10.3200/SOCP.148.5.631-636

VAN DIJK, W.W., & OUWERKERK, J.W. (2014). Introduction to schadenfreude. In W.W. van Dijk, & J.W. Ouwerkerk (Eds.), *Schadenfreude: Understanding pleasure at the misfortune of others*. Cambridge University Press.

VAN DIJK, W.W., OUWERKERK, J.W., GOSLINGA, S., & NIEWEG, M. (2005). Deservingness and schadenfreude. *Cognition and Emotion*, *19* (6), 933-939. psycnet.apa.org/doi/10.1080/02699930541000066

VAN DIJK, W.W., WESSELING, Y.M., OUWERKERK, J.W., & VAN KONINGSBRUGGEN, G.M. (2010). *The self-enhancing properties of another's misfortune*. Unpublished manuscript, Leiden University, The Netherlands.

WALTZ, C.F., STRICKLAND, O.L., & LENZ, E.R. (2005). *Measurement in nursing and health research*. Springer.

WANG, S., LILIENFELD, S.O., & ROCHAT, P. (2019). Schadenfreude deconstructed and reconstructed: A tripartite motivational model. *New Ideas in Psychology*, *52*, 111. doi.org/10.1016/j.newideapsych.2018.09.002

APPENDIX

Final Tripartite Scadenfreude Scale items

نیچے دیے گئے بیانات کو غور سے پڑھیں اور بتا ئیں کہ نیچے دی گئی صورتحال میں آپ کس قدر خوشی اور اطمینان محسوس کرتے ہیں ۔ اپنے بارے میں مناسب رائے کے اظہار کیلیے نیچے دئیے گئے ممکنہ جوابات میں سے موزوں ترین جواب کا انتخاب کریں۔

ہمیشہ	اکثر	کبھی کبھار	کبھی کبھی	شاذو نادر	کبھی نہیں	بیانات	
						AGGRESSION SCHADENFREUDE	**1**
5	4	3		2	1	میرے لیے یہ باعث مسرت ہے کہ سما جی لحاظ سے میرے مخالفین مجھ سے کمتر ہیں۔	1
5	4	3		2	1	اگر ایک شخص کو میری ذات /قبیلہ /برا دری کے لوگ برا سمجھیں تو اس شخص کی ناکامی پر مجھے خوشی ہو تی ہے۔	2
5	4	3		2	1	اپنے ساتھیوں کو مرعوب کر نے کیلیے انکے دشمنوں پر کیچڑ اچھالنا میری پسندیدہ حکمت عملی ہے۔	3
5	4	3			1	اپنی ذات /قبیلہ /برا دری سے کم تر لوگوں کو کامیاب ہو تے دیکھ کر میرا دل کڑھتا ہے۔	4
5	4	3		2	1	مجھے ایسی فلمیں دیکھنا پسند ہیں جس میں سما جی فرق کو ظلم و زیادتی اور مار پیٹ کے طور پر دکھایا گیا ہو۔	5
5	4	3		2	1	مجھے اپنے مخالفین کی برائیاں اور اپنے ساتھیوں کی اچھا ئیاں بیان کرنا اچھا لگتا ہے	6
5	4	3		2	1	اپنے مخالفین کی کامیابی کا سن کر مجھے غصہ آتا ہے۔	7
5	4	3		2	1	ذہین اور کامیاب لوگوں کو ناکام دیکھ کر مجھے بہت مزا آتا ہے۔	8
5	4	3		2	1	دوسروں کو پھسل کر گرتے دیکھ کر میں اپنی ہنسی پر قابو نہیں رکھ سکتا/سکتی۔	9
5	4	3		2	1	مجھے ان لوگوں کا مذاق اڑا کر مزا آتا ہے جنہوں نے میری شکل /لباس کا مذاق اڑایا ہو۔	10
5	4	3		2	1	میرے لیے لوگوں کو عروج کے بعد رو بہ زوال دیکھنا قابل مسرت ہے۔	11
5	4	3		2	1	مجھے مشہور لوگوں (سیاستدان، گلو کار ، کھلاڑی وغیرہ) کی کامیابیوں کی بہ نسبت ان کی ناکامیوں کے بارے میں جاننا اچھا لگتا	12
						RIVALRY SCHADENFREUDE	**2**
5	4	3		2	1	مجھے اپنی ذات/قبیلہ /خاندان کے لوگوں سے اپنے مرتبے کا موازنہ کرتے رہنا اچھا لگتا ہے۔	13
5	4	3		2	1	لوگوں کو نقصان پہنچا کر اپنا فائدہ حاصل کرنے کا الگ ہی مزہ ہے۔	14

بیانات		کبھی نہیں	کبھی کبھار	اکثر	شاذو نادر کبھی	ہمیشہ
15	اپنا فائدہ حاصل کرنے کیلیے مجھے بر طرح کے حربے آزمانے میں خوشی ہوتی ہے	1	2	3	4	5
16	دوسروں پر اپنی برتری قائم کرنے کے لیے میں قانون کو ہاتھ میں لے کر ان پر غالب آنے کو ترجیح دوں گا/ دوں گی	1	2	3	4	5
17	جو لوگ مختلف حربوں کے ذریعے مجھ پر حاوی ہونا چاہیں مجھے ان کو مغلوب کرنے میں بہت تسکین ملتی ہے۔	1	2	3	4	5
18	مجھے ذاتی مفاد حاصل کرنے کیلیے دوسروں کا استحصال کرنا پسند ہے۔	1	2	3	4	5
19	جب میں کسی دوسرے کو غلطی کرتے ہوئے دیکھوں تو میں خود کو زیادہ عقل مند اور قابل انسان سمجھتا /سمجھتی ہوں۔	1	2	3	4	5
20	مجھے لگتا ہے کہ دوسروں کی کمزوریوں سے فائدہ اٹھاتے ہوے آگے بڑھنے اور ترقی کرنے کا اپنا ہی مزا ہے۔	1	2	3	4	5
21	کسی کا چھوٹا موٹا نقصان کرکے اپنے سماجی رتبہ میں اضافہ کرنے میں کوئی حرج نہیں۔	1	2	3	4	5
22	اگر میرے کسی عمل سے مخالف ذات/قبیلہ /خاندان کے فرد کو نا قابل تلافی نقصان پہنچ رہا ہو تو ایسا کام کرنے میں مجھے خوشی ہو گی۔	1	2	3	4	5
23	مجھے ذاتی مفاد کیلیے دوسروں کو نقصان پہنچانے میں دلیری اور خوشی کا احساس ہوتا ہے۔	1	2	3	4	5
24	میں ایسی محافل میں جانا پسند کرتا /کرتی ہوں جن میں میرے حلقہ احباب والے ایسے شخص کو برا بھلا کہیں جس سے میرا ذاتی بغض ہو۔	1	2	3	4	5
3	**JUSTICE SCHADENFREUDE**					
25	مجھے ایسے شخص کو بے یارو مدد گار دیکھ کر خوشی ہوتی ہے جو خود دوسروں کی مدد نہیں کرتا۔	1	2	3	4	5
26	جب دھوکہ دینے والا خود دھوکہ پاتا ہے تو مجھے طمانیت کا احساس ہوتا ہے۔	1	2	3	4	5
27	میں چاہتا /چاہتی ہوں کہ جسے دوسروں کی مشکلات کا احساس نہیں ہوتا اسے خود ان مصائب کا سامنا کرنا پڑے۔	1	2	3	4	5
28	جو شخص دوسروں کی عزت نہیں کرتا اسے بےعزت ہوتے دیکھ کر مجھے دلی تسکین ملتی ہے۔	1	2	3	4	5
29	وہ شخص جو خاندان کی روایات سے بغاوت کرے اس کے ساتھ برا ہوتے دیکھ کر اطمینان کا احساس ہوتا ہے۔	1	2	3	4	5
30	کسی کو غلطی کی سزا پاتے دیکھ کر مجھے سکون کا احساس ہوتا ہے۔	1	2	3	4	5

English translation of Tripartite Schadenfreude Scale items

1.   It is gratifying that my opponents are socially inferior to me.

2.   If a person is considered bad by members of my community or tribe, I feel happy about that person's failure.

3.   To impress my companions, my favorite strategy is to accuse or insult their enemies.

4.   Seeing people inferior to my community or tribe succeed, makes me upset.

5.   I like watching movies that show social differences as oppression, abuse, and violence.

6.   I enjoy talking about my opponents' faults and my companions' virtues.

7.   Hearing about my opponents' success makes me angry.

8.   Seeing smart and successful people fail gives me great pleasure.

9.   I cannot control my laughter when I see others slipping and falling

10. I enjoy making fun of people who have mocked my appearance or my clothes.

11. It gives me pleasure to see people fall from grace after their success.

12. I prefer learning about the failures of famous people (like politicians, singers, athletes, etc.) rather than their successes.

13. I like comparing my status with the people of my caste, tribe, or clan.

14. There is a unique pleasure in gaining benefits by harming others.

15. I enjoy trying every possible tactic to achieve my goals.

16. To establish my superiority over others, I prefer taking the law into my hands and overpowering them.

17. I find great satisfaction in defeating those who try to dominate me through various tactics.

18. I like exploiting others for personal gain.

19. When I see someone else making a mistake, I consider myself an intelligent and capable human being.

20. I believe that taking advantage of others' weaknesses to advance and succeed has its pleasure.

21. There is nothing wrong with causing someone a small loss to enhance your social status.

22. If any of my actions cause irreparable harm to an individual from a rival caste/tribe/family, I would be happy to do so.

23. I feel brave and happy when I harm others for personal gain.

24. I like attending gatherings where my circle of friends criticizes someone, against whom I have personal grudges.

25. I feel happy to see a person helpless, who does not help others.

26. I feel satisfaction when someone who deceives others, gets deceived.

27. I wish that the person, who does not care about others' troubles, may face the same difficulties.

28. I feel inner peace seeing a person being humiliated, who does not respect others.

29. I feel contentment when a person who rebels against family traditions faces bad consequences.

30. I feel a sense of calm when I see someone being punished for his mistakes.

Evaluation of a gamified intervention for eco-driving: A pilot study

Richard Stephens

School of Psychology, Keele University, United Kingdom

r.stephens@keele.ac.uk

⚲ ABSTRACT. L'eco-driving è uno stile di guida del veicolo che riduce il consumo di energia, mentre la gamification si riferisce all'uso di tecniche di gioco in contesti non di gioco per motivare il coinvolgimento dell'utente. In questo studio è stato valutato un intervento di eco-guida gamificato basato su un'app per smartphone. Gli effetti dell'utilizzo dell'app sui punteggi dei parametri di guida ecologica nell'arco di 3 settimane di guida sono stati confrontati in quattro condizioni: Controllo (C1); Educazione (C2), Gamification (E1); Gamification + Educazione (E2). Le esperienze degli utenti suggeriscono che il sistema di punteggio dell'app è stato percepito come affidabile ed equo e i dati emersi suggeriscono che un approccio gamificato che utilizza un'app per smartphone può incoraggiare efficacemente la guida ecologica, con benefici statisticamente significativi. Un intervento di questo tipo ha il vantaggio di essere relativamente economico, dato che si basa sulla tecnologia degli smartphone esistente e ampiamente disponibile ma si rendono necessarie ulteriori ricerche con un campione più ampio per verificare che gli effetti mostrati siano effettivamente affidabili.

⚲ SUMMARY. *Eco-driving is a style of vehicle driving that reduces energy consumption, while gamification refers to use of game techniques in non-game contexts to motivate user engagement. A gamified eco-driving intervention based on a smartphone app was evaluated. In a mixed methods approach, effects of using the app on eco-driving parameter scores over 3 weeks of driving were compared across four conditions: Control (C1), Education (C2), Gamification (E1), Gamification + Education (E2). Users were also invited to share their experiences in a focus group. Comparing across the four conditions, effect were null. However, combining the control and user groups showed a benefit of using the app for overall score, harsh braking score, speeding score and leaderboard position. User experiences suggest the app scoring system was perceived as reliable and fair. Suggestions for improvements included redesigning the leaderboard and balancing intrinsic with extrinsic rewards for longer term engagement. These data suggest a gamified approach using a smartphone app may effectively encourage eco-driving, with statistically significant benefits of regular reflection on eco-driving scores. Such an intervention has the advantage of being relatively economical, given that it is based on existing, widely available smartphone technology. Nevertheless, further research with a larger sample is required to verify that the effects shown are reliable, and to improve psychological understanding of gamification applied to eco-driving.*

Keywords: *Driving, Eco-driving, Gamification, User-experience, Speeding*

INTRODUCTION

Background

Eco-driving is a style of vehicle driving that reduces energy consumption, maximizing mileage per unit of energy consumed (Stillwater & Kurani, 2013). Elements include adherence to speed limits, accelerating and braking smoothly and avoiding over revving (Magaña & Muñoz-Organero, 2015) as well as vehicle maintenance, trip planning and vehicle choice (Stillwater & Kurani, 2013). Eco-driving may save up to 25% of fuel (Kamal, Mukai, Murata & Kawabe, 2011).

Literature review

Gamification refers to use of game techniques in non-game contexts to motivate user engagement and participation (Deterding, Dixon, Khaled & Nacke, 2011). Gamification introduces intrinsically motivating elements such as target scores, as opposed to extrinsic rewards like cash. There is a rich literature showing that gamification can benefit driving in general (e.g. Diewald, Möller, Roalter, Stockinger & Kranz, 2013). Gamification applied to eco-driving was recently reviewed by Stephens (2022). The review identified 39 separate studies, of which 13 comprised evaluation studies employing quantitative indices of eco-driving. These studies comprised both lab-based (simulator) and real-world driving scenarios. However, only three of these were found to be of acceptable quality for drawing meaningful conclusions. Recurring methodological problems included conflation of intrinsic rewards of gamification with extrinsic rewards and absence of inferential statistical analysis.

The evaluation studies reviewed and found to be of acceptable quality included Magaña and Muñoz-Organero (2015), who evaluated a smartphone app-based system in 36 drivers over 2,160 road trips in Spain. They found lowered fuel consumption in participants who used the app, which provided instant feedback on eco-driving scores, in-game achievements and social comparisons, compared with controls. Steinberger and colleagues (Steinberger, Schroeter, Foth & Johnson, 2017; Steinberger, Schroeter & Watling, 2017) evaluated their coastmaster system which encouraged minimal braking during transitions from higher to lower speeds. Within-subjects simulator studies with 32 and 24 male drivers showed that using the system reduced overall

speed, driving over the speed limit, improved anticipation. The review concluded that gamification offers promise as a method of encouraging eco-driving but its efficacy in that domain remains largely unsubstantiated (Stephens, 2022).

The present study

In the present study a smartphone app-based gamification system relevant to eco-driving was evaluated. The *Safest Driver* smartphone app (Cambridge Mobile Telematics) was chosen as it had the most functionality relevant to gamifying eco-driving compared with similar widely available apps, while having extensive compatibility and free availability on the iOS (Apple) and Android operating systems. This choice recognized that safe/eco-driving overlap, with both driving styles sharing goals including reducing average speed and avoiding harsh acceleration and braking (Vaezipour, Rakotonirainy, Haworth & Delhomme, 2019).

Based on GPS and other data, the *Safest Driver App* compiles daily scores out of 100 for: distraction (mobile phone usage), speeding (exceeding posted speed limits), braking (braking harshly), acceleration (accelerating harshly), cornering (excessive G-force). An overall Safe Driving score combining these was used as a surrogate indicator of eco-driving. A leaderboard also updates daily. As this app runs in the background it presented no driving safety hazard due to mobile phone distraction.

As mentioned already, one aim of gamification is to generate intrinsic motivation. Certain elements of intrinsic motivation, specifically, satisfying work with clear goals and tasks, and hopes/experiences of success (McGonigal, 2011) overlap with psychological flow theory (Csikszentmihalyi & LeFevre, 1989; Šimleša, Guegan, Blanchard, Tarpin-Bernard & Buisine, 2018). Psychological flow is experienced when there is a good match between the challenges presented by a situation and the skills a person possesses to meet such challenges. In the context of eco-driving, gamification may generate flow and consequent intrinsic motivation and enjoyment by increasing the level of challenge of the otherwise mundane task of driving. Therefore, flow, or relatedly enjoyment, may mediate effects of gamification on eco-driving. As this has not been investigated previously in the context of gamification applied to eco-driving, measures of flow and enjoyment were included.

Sensation seeking, defined as "the seeking of varied, novel, complex, and intense sensations and experiences and the willingness to take physical, social, legal, and financial risks for the sake of such experiences" (Zuckerman, 1994, as cited by Jonah, 1997), was assessed as a possible moderating variable. Fewer indications of eco-driving were predicted for individuals with higher sensation seeking scores. Acceptance of technology was also assessed as a possible moderator, predicting fewer indications of eco-driving for individuals less accepting of technology. Finally, eco-driving knowledge was assessed as a further moderator, predicting fewer indications of eco-driving for more knowledgeable individuals already performing eco-driving behaviours ahead of the intervention.

A mixed methods approach was applied. A quantitative phase lasted three weeks with four conditions: Control: asked to drive normally (C1); Education: reviewed an eco-driving information package and challenged to adopt an eco-driving style (C2); Gamification: used the *Safest Driver* app checking scores daily and challenged to adopt an eco-driving style (E1); Gamification + Education: used the *Safest Driver App* checking scores daily and challenged to adopt an eco-driving style, plus reviewed an eco-driving information package (E2). These conditions were in place over weeks 1 and 2, but in week 3 all participants used the *Safest Driver App* checking scores daily and were challenged to adopt an eco-driving style. This meant all participants could share experiences of using the app in user-experience focus groups.

Hypotheses

(i) Eco-driving parameters (overall score; speeding score; braking score; acceleration score; leaderboard position) will be higher in condition E2 *vs* E1, in condition C2 *vs* C1, and in the combined experimental groups (E1 and E2) compared with the combined control groups (C1 and C2).

(ii) Flow and enjoyment will be higher for E2 *vs* E1, for C2 *vs* C1, and for combined experimental groups v combined control groups.

(iii) Changes in eco-driving parameters will be mediated by psychological flow and/ or enjoyment.

(iv) Effects assessed under hypotheses (i) and (ii) will be moderated by sensation seeking, acceptance of technology, and eco-driving knowledge.

METHOD

Participants

Recruitment adverts were placed physically around a university campus and on Twitter. Prerequisites were: verified informed consent, UK based, aged 18+, full UK driving licence, drive regularly (over two hours weekly), access to an insured, road-legal vehicle. The final sample ($N = 24$) comprised 15 females, 8 males and 1 non-binary individual of mean age 33.5 years ($SD = 13.3$). Compensation was £8 Amazon vouchers per hour for up to 9-hours (total £72.00). This sample size was chosen based on affordability within the available resources for carrying out the study. The study protocol was reviewed favourably by the Keele University Research Ethics Committee.

Design

Quantitative data were analyses in a 3×4 mixed design with participants randomly allocated to one of 4 conditions: Control (C1), Education (C2), Gamification (E1), Gamification + Education (E2). The within-subjects variable was week (1, 2, 3). Dependent variables were relevant app scores (overall score, speeding score, braking score, acceleration score, leaderboard position), Engeser and Ulrich flow scale scores, enjoyment ratings. Questionnaire scores were employed as possible mediator or moderator variables. A qualitative phase gathered user experiences via a series of focus groups guided by a question script.

Materials

– *Safest Driver App (Cambridge Mobile Telematics)*. This outputs daily 0-100 scores for distraction, speeding, braking, acceleration and cornering. These are combined into an overall score; a higher value is consistent with eco-driving. Distraction and cornering scores, irrelevant to eco-driving, were discarded.
– *Safest Driver User Guide*. A five-page illustrated guide to the functionality of the *Safest Driver App* was created.
– *Eco-driving Education Package*. This was assembled based on eco-driving advice published by the UK Energy Saving Trust (https://energysavingtrust.org.uk/advice/efficient-driving/).

– *Driving context*. Participants estimated the proportion of driving (a) built-up areas with 20/30 mph limits; (b) A or B-roads with 40-60 mph limits; (c) dual carriageways/ motorways. They also provided make, model, year, fuel type, engine size (where applicable) of their car, and whether anyone else had driven it.

– *Enjoyment Visual Analogue Scale* (*VAS*). Driving enjoyment ("I enjoyed driving over the last week") was rated on a horizontal line anchored left, "Not at all" and right, "Very much" using a graphic slider scored 0-100, as used by Stephens and Smith (2022).

– *Questionnaires*. These were the *Engeser Short Flow Scale* (10-items; Engeser & Baumann, 2016); the *Flow Index* (3-items; Ulrich, Keller, Hoenig, Waller & Grön, 2014); the *Brief Sensation Seeking Scale* (8-items; Hoyle, Stephenson, Palmgreen, Lorch & Donohew, 2002); the *Acceptance of Technology Scale* (9-items; Van der Laan, Heino & De Waard, 1997); the *Eco-driving Knowledge Scale* (5-items; Günther, Kacperski & Krems, 2020).

Procedure

Participants attended an induction meeting via video call. After providing informed consent, a baseline survey collected demographic information (age in whole years; gender with the options: female, male, non-binary, prefer not to say); years of holding a full driving licence; usual weekly driving time; understanding of eco-driving (open text response); understanding of strategies to achieve eco-driving (open text response); extent of practicing eco-driving (open text response). Participants next completed the questionnaires.

After installing the *Safest Drive*r smartphone app on their phone, participants were randomised to one of the four conditions: Control (C1); Education (C2); Gamification (E1); Gamification + Education (E2). The E1 and E2 groups were briefed on app functionality. The C2 and E2 Education groups were given several minutes to study the *Eco-driving Education Package*. The C1 group was instructed: "When you are out in the car, please drive the way you would normally". Alternatively, C2, E1 and E2 were instructed: "We challenge you to try and become more of an eco-driver, that is, to reduce how much energy you use during driving". C1 and C2 were further directed not to open the *Safest Driver App* for the first 2 weeks of driving. Participants received copies of all instructions and information. E1 and E2 were prompted

each evening to enter app scores into an online survey. The instructions changed for C1 and C2 at week 3. They were sent the *Safest Driver User Guide*, challenged to try and become an eco-driver, and prompted daily to check and enter their app scores into a survey.

Weekly online surveys asked participants to complete the flow scales, the enjoyment VAS and driving context information for driving over the last week. Current average mpg (or miles per KWh) from their car's computer display was also requested where available.

User experiences were shared by 23 participants in one of four, 1-hour, online focus groups. A question script guided discussion for the topics: knowledge about eco-driving; general usability of the app; specific app features; gamification; extrinsic rewards; suggested additional features; study procedures; open comments. Discussions were transcribed for thematic analysis, which aimed to detect pertinent patterns in the data. These themes were then developed based on how users talked about their experiences of using the gamified app, with the aim being interpretation and sense-making of the data.

RESULTS

Quantitative data analysis

Due to the small sample size, no effort was made to manage outliers. Skewness and Kurtosis coefficients were within an acceptable range except overall score in week 2 (Skewness = −1.530, Kurtosis = 3.600), and speeding scores across weeks 1-3 (Skewness −2.456 to −2.244, Kurtosis 4.488 to 6.239). Therefore, a mixture of parametric and non-parametric analyses were applied. Descriptive data are shown in Table 1.

Hypothesis (i) was assessed in a series of 3×4 mixed ANOVAs (week: 1, 2, 3 × condition: C1, C2, E1, E2) for the dependent variables: overall score; braking score; acceleration score; and leaderboard position. All main and interaction effects were null ($p<.05$), apart from the main effect of week on leaderboard position, $F_{(2,40)} = 9.951$, $p<.001$, $\eta^2 p = .332$. Holm-corrected contrasts showed lower (i.e. superior) leaderboard position in week 2 *vs* 1, $p<.001$, and in week 3 *vs* 1, = .002. The main effect of condition (C1, C2, E1, E2) on Speed was assessed using a Kruskal-Wallis test, which found no effect ($p>.05$). The main effect of week (1,2,3) on speed was

Table 1 – Descriptive data of a combination of parametric and non-parametric analyses

	Control (C1)	Education (C2)	Gamification (E1)	Gamification + Education (E2)
Age	38.2 (13.4)	26.8 (12.4)	35.7 (11.8)	33.3 (16.0)
Gender[1]	3/3/0	3/2/1	5/1/0	4/2/0
Sensation seeking	2.50 (.60)	3.50 (.65)	2.88 (1.11)	2.46 (.48)
AOT (usefulness)	2.30 (.17)	2.60 (.61)	2.90 (.72)	2.23 (.32)
AOT (satisfaction)	3.67 (.20)	3.13 (.21)	3.13 (.26)	3.00 (.32)
Eco-driving knowledge	4.90 (1.67)	4.90 (1.37)	3.97 (1.69)	4.80 (1.50)
Licence years	19.00 (13.19)	7.21 (12.32)	14.17 (8.95)	14.83 (14.28)
Engine cc of main car[2]	1,652 (321)	1,187 (135)	1,200 (236)	1,567 (234)
Number of scored trips[3]	81.0 (23.0)	91.5 (26.1)	77.8 (55.3)	56.7 (19.4)
Number of scored km[3]	1,006 (386)	1,135 (449)	590 (389)	824 (543)
Fuel type[2,4]	2/3/0/1	6/0/0/0	4/2/0/0	4/2/0/0
Ratio of road types[2,5]	52/38/10	50/38/12	66/26/8	40/34/26
Passenger[2,6]	4	2	5	5

Note. [1] Frequencies for female/male/non-binary; [2] Recorded after week one of driving; [3] Recorded across all three weeks of driving; [4] Frequencies for petrol/diesel/hybrid/BEV; [5] Mean percentage for 20-30mph/40-60mph/70mph road types; [6] Frequencies for passengers in another car at least once.

assessed using a Friedman test, which also found no effect ($p > .05$). These data are illustrated in Figure 1.

Acknowledging the low power afforded by the small sample, the two experimental groups were combined into a single group (E1 and E2), as were the two control groups (C1 and C2). A 3×2 ANOVA showed a main effect of condition for overall score, $F_{(1,22)} = 6.631$, $p = .017$, $\eta^2 p = .232$. As the interaction was close to significant, $F_{(2,44)} = 2.778$, $p = .073$, $\eta^2 p = .112$, contrasts were computed. These showed higher overall scores in the experimental *vs* the control group in week 1, $p = .005$ and week 2, $p = .025$, but not week 3, $p = .051$.

The condition × week interaction was significant for

Figure 1 – Overall score (a), speeding score (b), braking score (c), acceleration score (d), and leaderboard rank (e) across the three weeks of driving (weeks 1, 2 and 3), by experimental group
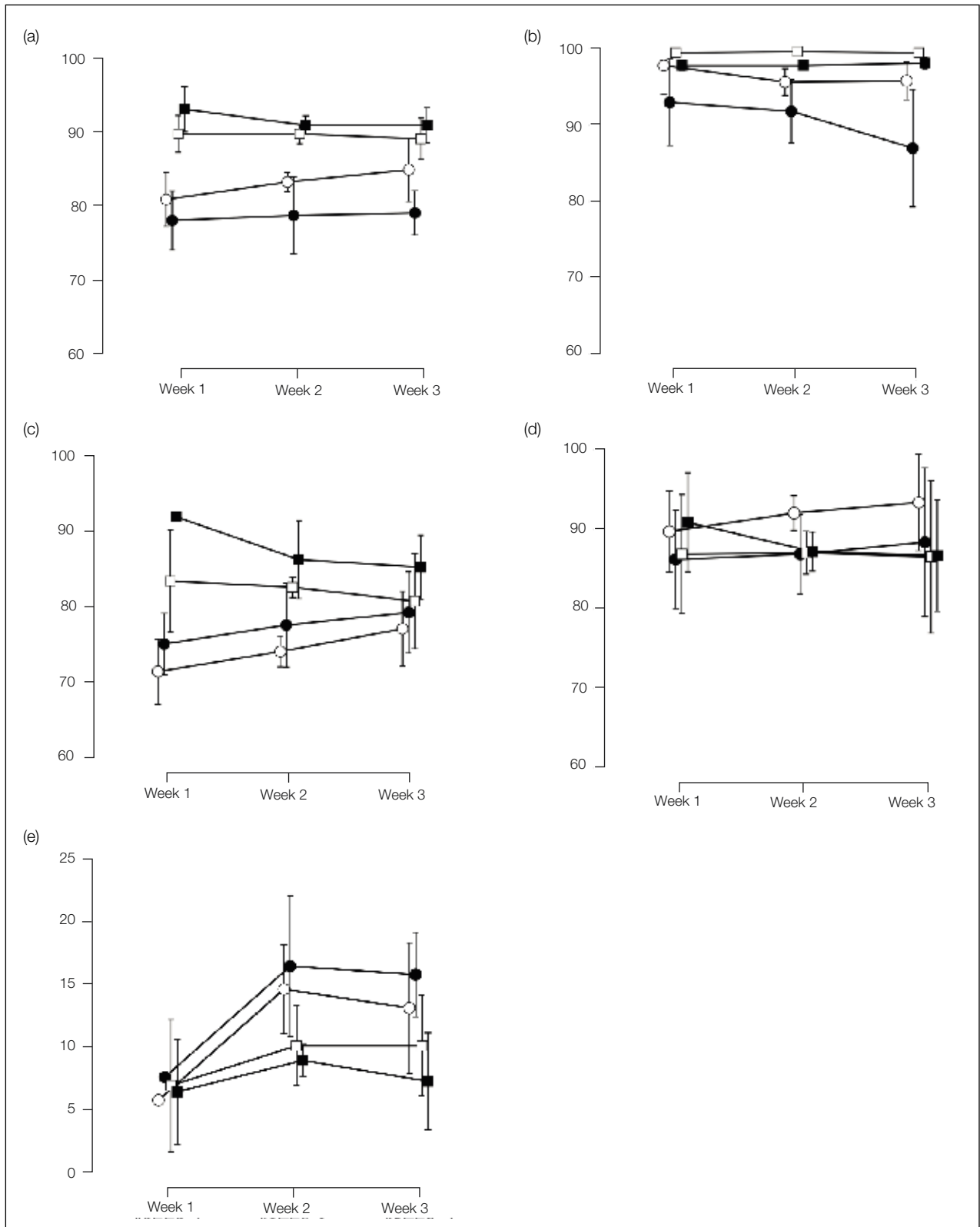
Legenda. White circle = C1; dark circle = C2; white square = E1; dark square = E2.

braking, $F_{(2,44)} = 5.912$, $p = .005$, $\eta^2 p = .212$. Contrasts showed higher braking scores in the experimental *vs* the control group in week 1, $p = .003$, but not week 2, $p = .065$, or week 3, $p = .289$. Main and interaction effects were null for acceleration score ($p>.05$). For leaderboard position, the interaction was close to significant, $F_{(2,44)} = 3.188$, $p = .051$, $\eta^2 p = .127$, and contrasts were computed. These showed lower (i.e. superior) leaderboard position in the experimental group in week 2, $p = .026$, and week 3, $p = .032$, but not week 1, $p = .999$.

Speeding was assessed using non-parametric tests, showing a main effect of condition, $F_{(1,22)} = 5.607$, $p = .027$, $\eta^2 p = .203$. Mann-Whitney tests showed higher scores in the experimental group in week 2, $p = .009$, and week 3, $p = .033$, but not week 1, $p = .115$. Overall, these analyses support hypothesis 1, with evidence that overall score, speeding score, braking score and leaderboard position differed across the experimental and control groups in a direction consistent with increased eco-driving. These data are illustrated in Figure 2.

Hypothesis (ii) analyses first compared across conditions C1, C2, E1, E2, and then across the combined experimental (E1 and E2) and control (C1 and C2) groups. Across four conditions all main and interaction effects were null ($p>.05$), apart from the main effect of week on enjoyment rating, $F(2,28) = 3.476$, $p = .045$, $\eta^2 p = .199$. Holm-corrected contrasts showed higher enjoyment ratings for week 3 *vs* 2, $p = .044$, but no effect for week 2 *vs* 1, $p = .204$, or week 3 *vs* 1, $p = 372$. For the combined experimental and control groups all main and interaction effects were also null ($p>.05$). Overall, this hypothesis was not supported. Descriptive data are shown in Table 2.

Hypothesis (iii) analyses began with checking correlations between the overall eco-driving scores, the two flow scale scores and the enjoyment rating scores across the three driving weeks. As none of these correlations were significant ($p>.05$) this hypothesis was not supported.

Hypothesis (iv) was assessed in analyses of covariance including condition, combined experimental groups (E1 and E2) compared with combined control groups (C1 and C2), week (1, 2, 3) and one of the covariates: sensation seeking, acceptance of technology (usefulness), acceptance of technology (satisfaction) or eco-driving knowledge. The dependent variables were: overall score, braking, speeding, leaderboard position. As none of the three-way interactions were significant, $F_{(2,40)}<1.65$, $p>.205$, this hypothesis also was not supported.

Qualitative data analysis

Thematic analysis (Braun & Clarke, 2006) of focus group transcripts was carried out solely by the author. The decision to employ one coder was taken in view of the limited scope of this pilot study. Limitations of not asking a second person to code a sample of the data are discussed in the limitations section of the Discussion. The coding process comprised several stages. First, the author reviewed the transcripts and noted initial codes. Then, initial codes were refined to identify overarching elements and sub-themes. A third stage involved selecting quotes that aligned with these themes, followed by a review and naming of the themes. Once finalized, the report writing commenced. The codebook is included as supplementary material in Appendix. The analysis identified seven themes: Intrinsic motivation, Eco-driving, App positives, App niggles, Real world context, Wider concerns, It's just not for me. These are described below.

Intrinsic motivation. Gamification harnesses intrinsic motivation towards promoting desirable behaviours such that any scoring system must be perceived as fair. This was indeed the perception of participants, e.g. "I think the metrics on the whole were pretty good. And they, certainly from an eco-perspective, acceleration and braking, were probably pretty important aspects of that (08)"; "When I was in the car with other people who aren't quite as safe drivers. The score was going down. So yeah, I think it was really accurate (26)". Participants enjoyed using the app, e.g. "I did find it fun and I was telling my family about it (26)"; "It made me think about things I hadn't really thought about, but in a way that made it quite fun (12)".

Eco-driving. Participants discussed specific improvements in eco-driving technique, including accelerating more gently, e.g. "I didn't realize how harsh I accelerate until I started using the app (21)"; "I definitely watch like my rev counter more now (19)", avoiding sudden braking, e.g. "I think my braking score went up once I was seeing the app and I was kind of more aware of that (26)"; "So, then, that would say to me, yeah, probably I'm too close to the ones in front (23)", and speeding less, e.g. "I think it encouraged me to think like, ohh, I am actually sticking to the speed limit and doing pretty well (12)".

App positives. Participants experienced the app as user-friendly, e.g. "It was very straightforward and it was very appealing to look at (24)". Some liked the mapping feature, e.g. "It was quite nice to see where I'd been (17)"; "I'm used

Figure 2 – Overall score (a), speeding score (b), braking score (c), acceleration score (d), and leaderboard rank (e) across the three weeks of driving (weeks 1, 2 and 3), by experimental group
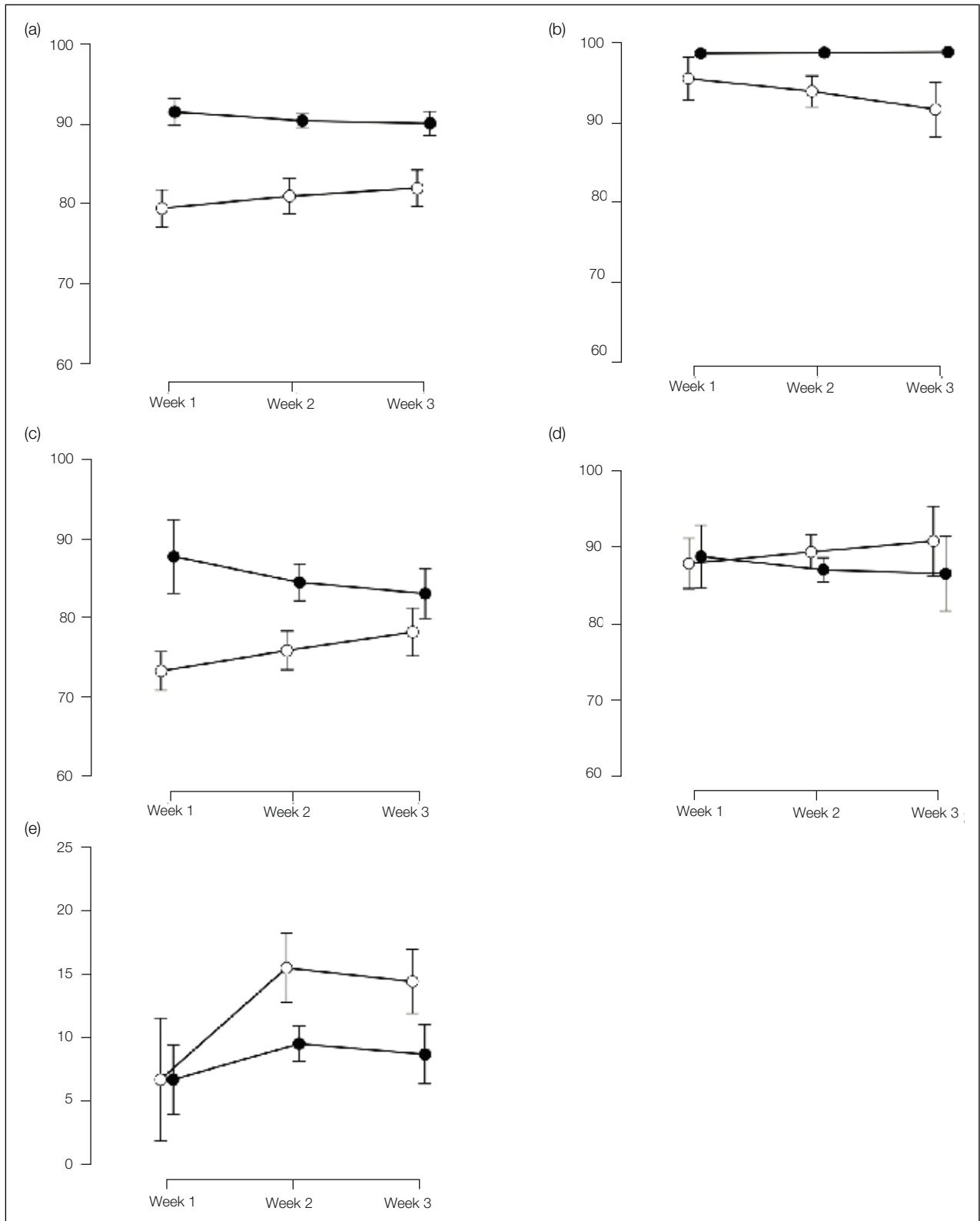
Legenda. White circle = combined control group; dark circle = combined experimental group.

Table 2 – Means (*SD*s) for the Engeser and Ulrich flow scales, and enjoyment rating across weeks 1-3 of driving, by condition

		Control (C1)	Education (C2)	Gamification (E1)	Gamification + Information (E2)
Engeser flow	Week 1	5.37 (.55)	5.05 (.50)	4.68 (.45)	5.13 (.63)
	Week 2	5.47 (.55)	4.10 (1.13)	5.03 (.77)	5.13 (.54)
	Week 3	5.33 (.67)	5.10 (.71)	5.08 (.83)	5.32 (.52)
Ulrich flow	Week 1	15.33 (3.08)	15.67 (2.94)	18.00 (3.16)	16.00 (2.53)
	Week 2	15.33 (3.20)	14.50 (3.02)	16.00 (3.10)	14.50 (1.76)
	Week 3	14.50 (7.45)	11.17 (8.73)	13.67 (7.45)	14.17 (7.28)
Enjoyment	Week 1	73.00 (28.41)	80.67 (15.63)	80.67 (16.37)	83.00 (12.19)
	Week 2	71.00 (28.73)	61.00 (40.04)	78.50 (21.55)	80.50 (17.92)
	Week 3	81.20 (26.86)	87.00 (16.09)	80.50 (23.56)	82.75 (4.92)

to… running apps and stuff like that... so it felt like the most visually, like familiar in terms of the app (12)". Participants described the leaderboard enhancing motivation, e.g. "I wanted to win and beat other people's scores. So yeah, I think it did encourage me slightly (14)"; "I think the competition element was quite healthy (16)". The driving tips feature was also welcomed, e.g. "I got tips and yeah, I did take them on board and I think it did increase sort of the scores (26)".

App niggles. Numerous participants expressed frustration when emergency braking impacted their score, e.g. "It was either that or mow down the pedestrian in front of me (11)". Some found the driving tips menus repetitive, e.g. "It was just the same thing every time… and it was just constantly - Ohh try not to accelerate. Ohh try not to brake so hard (11)"; "They didn't tell you, like, how to improve it (14)". Discussing absent features several would have welcomed specific challenges, like improving braking score: "Having short term challenges to, especially if there's a reward or something for it, then yeah, definitely (08)". One participant wished for more integration

with social media: "If you could post your scoreboards straight to your socials or something like that might be a way of competing with friends (08)". Suggestions for extrinsic rewards were numerous, including reduced insurance costs, charitable donations, shopping vouchers and fuel discounts, e.g. "For every 10,000 points you get will donate even if it's 10p, you know, to some kind of you know, sustainability or whatever (17)". Another suggestion was tangible feedback on fuel savings, e.g. "If the 'overall score' had a 'you've saved ××× amount on petrol this week compared to the average driver' (12)". Introducing live feedback received a mixed evaluation. Participants balanced benefits, e.g. "Accelerating or, like, going over the speed limit, I think that's fair enough for it to ping, just for your own safety and others (06)", with potential for annoyance or distraction, e.g. "I think I find a distraction. I just wanna swear at it (17)".

Real world context. Tension was expressed between eco-driving at the speed limit and other road users speeding, e.g. "I would be irritating the person behind me who was trying

to get home faster than me (09)". Tension was also expressed between scores and genuine eco-driving, e.g. "You've got to put your foot down a little bit to actually stay within eco driving, otherwise you're sitting there for 15-20 minutes, wasting all that fuel just sitting there (05)"; "If you were driving, which… shouldn't be driving 50 in a 30, it would actually probably be more eco but it wouldn't be safe (08)". Several participants found their enthusiasm for using the app plateaued, e.g. "It became quite a chore sort of looking at it every day because I come home, got to get the tea on and it's, it's another thing I've got to do (09)".

Wider concerns. One privacy concern equated the app with insurance tracking devices, e.g. "I know personally if, if I saw an offer to have a black box I would 100% not choose that (16)". Another was concern over traffic infringements being reported: "How fast could you go without them notifying the authorities? (08)". There was concern over the limits of gamification for behaviour change, e.g. "If people wanted to treat it like a game, so they have to find some way to manipulate the data to make them seem like they're driving better than they actually are (16)". A further concern was the bigger picture around pro-eco behaviours beyond eco-driving: "We've recently moved house and positioned ourselves so that I can walk to work. I can walk my kids to nursery, I can walk them to school (11)".

It's just not for me. Finally, several drivers did not like the app, e.g. "For me personally, I don't think I really changed how I drove throughout it (24)"; "It didn't really make me alter my ways (04)". One issue was perceived interference with the sense of freedom that driving can provide, e.g. "It took away, it made it harder for me to drive (05)".

DISCUSSION

Results in context

This pilot study evaluated a gamified approach to encouraging eco-driving based on a smartphone app. A quantitative data analysis comparing users and non-users of an eco-driving app was followed by a qualitative analysis of user experiences.

The configuration of the study with four conditions (C1, C2, E1, E2) lacked the statistical power to detect any differences across conditions. However, comparing the combined experimental groups (E1 and E2) with the combined control groups (C1 and C2) enabled, over two weeks of driving, comparisons between an experimental group of individuals reflecting daily on their Safest Driver app scores and trying to improve them ($n = 12$), with a control group of individuals driving with the app running in the background without reviewing their scores ($n = 12$). Hypothesis (i) that the eco-driving parameters assessed by the Safest Driver app would show improvement in the experimental conditions (E1 and E2) over the control conditions (C1 and C2), was supported with evidence of increased overall score, braking score, speeding score and higher leader board position in the experimental group. In demonstrating reduced incidences of harsh braking and speeding, these data suggest tangible benefits of gamified approaches to eco driving, at least in the short-term. This finding is in line with the findings of a recent review of gamification applied to eco-driving (Stephens, 2022). However, a larger-scale study with consequent increased statistical power is required to verify these effects.

The psychological mechanism underlying these gamification effects remains unclear.

While hypothesis (ii) was unsupported, with no effects for psychological flow or enjoyment, there was insufficient data to eliminate these mechanisms. Further research should assess this hypothesis with greater statistical power. Hypotheses (iii-iv) predicting mediation and moderation effects could also not be adequately tested due to low statistical power. This limitation should be addressed in a higher powered study.

The qualitative data indicate that, in delivering a trustworthy scoring system, the Safest Driver app was fit for purpose. Users were able to identify and reflect upon specific eco-driving techniques, including accelerating and braking more gently and avoiding speeding. They liked both the trips menu which displayed their prior journeys as a trace on a map with key incidents flagged and the driving tips, although more sophisticated and detailed tips would have been desirable. The leaderboard aspect received mixed reviews, consistent with previous studies (e.g. Stephens, 2022; Vaezipour, Rakotonirainy & Haworth, 2016). There was an appetite for personalised leader boards enabling competition against known other people such as family, friends or work colleagues, perhaps via social media.

There was, however, a plateauing of enthusiasm for using the app, suggesting a solely gamified approach to eco driving via intrinsic motivation may be time-limited, as has been suggested by Rapp and Boldi (2023) in their

study assessing the lived and meaning-laden experience of behaviour change. One reason for this may have been the reported tension between eco-driving and real-world driving, such as perceptions of holding up other drivers and being slow pulling out at junctions. Such concerns are known to influence road user behaviour (McNabb, Kuzel & Gray, 2017) and additional motivation from extrinsic rewards such as savings on fuel, car insurance, driving-related gadgets or other rewards may help to prolong the period of engagement, although further research would be required to assess this. A further cause of plateauing may have been privacy concerns. To counter this, specific reassurances could be made that data will not be shared with insurance companies or law-enforcement authorities. Some users may also benefit from reassurances that steps are being taken to prevent unintended negative consequences of gamification that undermine eco-driving. Perhaps a user-reporting mechanism could be put in place where such concerns could be raised. Some users did not find the app useful at all. While good design may win over a certain percentage of reluctant users, a proportion of individuals may be unwilling to engage with an eco-driving app. This may reflect the habitual nature of driving in which individual preferences forged over extended time periods become resistant to change (Caraban, Karapanos, Gonçalves & Campos, 2019).

Limitations

This small-scale pilot study had several limitations. A key issue was the small sample size, as already mentioned. While the study indicated beneficial effects of a gamified app for several eco-driving parameters (overall score, braking score, speeding score and higher leader board position), these effects could be artefacts of low statistical power. Consequently, they should be treated with caution until such time as they are verified in a larger-scale study with consequent increased statistical power. Further, the sample was a convenience sample which limits generalisability. Relatedly, the limited data collected should be considered of low reliability and validity. In interpreting the data the reader should bear in mind that this was a pilot study.

A further issue was the absence of an independent measure of eco driving out-with app scoring. An independent measure would show whether eco-driving was genuinely improving, rather than scores on an app which, though related to eco driving, may capture something else. Participants were asked to report weekly mpg readings from on-board car computer displays but take-up was low, probably because relatively few cars have this feature. Such independent eco-driving scores may be obtained in future from vehicle on-board computers or via customer fuel purchase data alongside present vehicle mileage.

In addition, the thematic analysis lacked investigator triangulation due to the absence of a second data coder. While triangulation has long been known to offer a solution to overcoming individual bias on the part of an investigator (e.g. Campbell & Fiske, 1959) no such undertaking was employed in the present study. This was a consequence of its status as a pilot study designed to trial methods and measures with limited resources. As with the other results in this study, the qualitative findings should be viewed with caution, and the author recommends employing investigator triangulation in future studies applying thematic analysis to gamified approaches to encouraging eco-driving. Further insights may also have been gained with recourse to established qualitative methodologies specific to human computer interaction such as heuristic evaluation and could usefully be explored in future studies.

CONCLUSION

In conclusion, this study finds that a gamified approach to encouraging eco-driving has potential to impact behaviour. A small-scale quantitative evaluation explored statistically significant benefits of regular reflection on scored aspects of eco-driving provided by a smartphone app, specifically, reducing harsh braking and speeding. User experiences reflected a general acceptance of the app including reflection upon specific aspects of eco-driving technique which the app helped improve. Such an intervention has the advantage of being relatively economical, given the wide availability of smartphone technology. Further confirmatory research should optimise study power and balance intrinsic and extrinsic rewards to promote prolonged engagement.

Competing interests: The author has no competing interests to declare.

References

BRAUN, V., & CLARKE, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3* (2), 77-101. doi.org/10.1191/1478088706qp063oa

CAMBRIDGE MOBILE TELEMATICS (2022). *Safest Driver*. Accessed 8th July 2022 from https://play.google.com/store/apps/details?id=com.safestdriver.drivingapp

CAMPBELL, D.T., & FISKE, D.W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56* (2), 81-105. doi.org/10.1037/h0046016

CARABAN, A., KARAPANOS, E., GONÇALVES, D., & CAMPOS, P. (2019). 23 ways to nudge. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. doi.org/10.1145/3290605.3300733

CSIKSZENTMIHALYI, M., & LEFEVRE, J. (1989). Optimal experience in work and leisure. *Journal of Personality and Social Psychology*, *56* (5), 815-822. doi.org/10.1037/0022-3514.56.5.815

DETERDING, S., DIXON, D., KHALED R., & NACKE, L. (2011). From game design elements to gamefulness: Defining gamification. *Proceedings of the 15th International Academic MindTrek Conference: Envisioning Future Media Environments*. doi.org/10.1145/2181037.2181040

DIEWALD, S., MÖLLER, A., ROALTER, L., STOCKINGER, T. & KRANZ., M. (2013). Gameful design in the automotive domain: Review, outlook and challenges. *AutomotiveUI '13: Proceedings of the 5th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*. doi.org/10.1145/2516540.2516575

ENGESER, S., & BAUMANN, N. (2016). Fluctuation of flow and affect in everyday life: A second look at the paradox of work. *Journal of Happiness Studies*, *17* (1), 105-124. doi.org/10.1007/s10902-014-9586-4

GÜNTHER, M., KACPERSKI, C., & KREMS, J.F. (2020). Can electric vehicle drivers be persuaded to eco-drive? A field study of feedback, gamification and financial rewards in Germany. *Energy Research & Social Science*, *63*, 101407. doi.org/10.1016/j.erss.2019.101407

HOYLE, R., STEPHENSON, M., PALMGREEN, P., LORCH, E., & DONOHEW, R. (2002). Reliability and validity of a brief measure of sensation seeking. *Personality and Individual Differences*, *32* (3), 401-414. doi.org/10.1016/s0191-8869(01)00032-0

JONAH, B. (1997). Sensation seeking and risky driving: a review and synthesis of the literature. *Accident Analysis & Prevention*, *29* (5), 651-665. doi.org/10.1016/s0001-4575(97)00017-1

KAMAL, M.A.S., MUKAI, M., MURATA, J., & KAWABE, T. (2011). Ecological vehicle control on roads with up-down slopes. *IEEE Transactions on Intelligent Transportation Systems*, *12* (3), 783-794. doi.org/10.1109/TITS.2011.2112648

MAGAÑA, V.C., & MUÑOZ-ORGANERO, M. (2015). GAFU: Using a gamification tool to save fuel. *IEEE Transactions on Intelligent Transportation Systems*, *7* (2), 58-70. doi.org/10.1109/MITS.2015.2408152

McGONIGAL, J. (2011). *Reality is broken: Why games make us better and how they can change the world*. London: The Penguin Group.

McNABB, J., KUZEL, M., & GRAY, R. (2017). I'll show you the way: Risky driver behavior when "following a friend". *Frontiers in Psychology*, *8*. doi.org/10.3389/fpsyg.2017.00705

RAPP, A. & BOLDI, A. (2023). Exploring the lived experience of behavior change technologies: Towards an existential model of behavior change for HCI. *ACM Transactions on Computer-Human Interaction*, *30* (6), Article 81. doi.org/10.1145/3603497

ŠIMLEŠA, M., GUEGAN, J., BLANCHARD, E., TARPIN-BERNARD, F., & BUISINE, S. (2018). The flow engine framework: A cognitive model of optimal human experience. *Europe's Journal of Psychology*, *14* (1), 232-253. doi.org/10.5964/ejop.v14i1.1370

STEINBERGER, F., SCHROETER, R., FOTH, M., & JOHNSON, D. (2017). Designing gamified applications that make safe driving more engaging. *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*, 2826-2839. doi.org/10.1145/3025453.3025511

STEINBERGER, F., SCHROETER, R., & WATLING, C.N. (2017). From road distraction to safe driving: Evaluating the effects of boredom and gamification on driving behaviour, physiological arousal, and subjective experience. *Computers in Human Behavior*, *75*, 714-726. doi.org/10.1016/j.chb.2017.06.019

STEPHENS, R. (2022). A review of gamified approaches to encouraging eco-driving. *Frontiers in Psychology*, *13*. doi.org/10.3389/fpsyg.2022.970851

STEPHENS, R., & SMITH, M. (2022). Effect of speed on flow and enjoyment for driving and rollercoasters. TRF: *Traffic Psychology and Behaviour*, *85*, 276-286. doi.org/10.1016/j.trf.2022.02.001

STILLWATER, T., & KURANI, K.S. (2013). Drivers discuss ecodriving feedback: Goal setting, framing, and anchoring motivate new behaviors. *Transportation Research Part F: Traffic Psychology and Behaviour*, *19*, 85-96. doi.org/10.1016/j.trf.2013.03.007

ULRICH, M., KELLER, J., HOENIG, K., WALLER, C., & GRÖN, G. (2014). Neural correlates of experimentally induced flow experiences. *Neuroimage*, *86*, 194-202. doi.org/10.1016/j.neuroimage.2013.08.019

VAEZIPOUR, A., RAKOTONIRAINY, A., & HAWORTH, N. (2016). Design of a gamified interface to improve fuel efficiency and safe driving. In *Design, user experience, and usability: Novel user experiences.* Cham: Springer. doi.org/10.1007/978-3-319-40355-7_31

VAEZIPOUR, A., RAKOTONIRAINY, A., HAWORTH, N., & DELHOMME, P. (2019). A simulator study of the effect of incentive on adoption and effectiveness of an in-vehicle human machine interface. *Transportation Research Part F: Traffic Psychology and Behaviour*, *60*, 383-398. doi.org/10.1016/j.trf.2018.10.030

VAN DER LAAN, J.D., HEINO, A., DE WAARD, D. (1997). A simple procedure for the assessment of acceptance of advanced transport telematics. *Transportation Research Part C: Emerging Technologies*, *5*, 1-10. doi.org/10.1016/S0968-090X(96)00025-3

APPENDIX

Codebook for gamification/eco-driving pilot study (March 2024)

Themes	Codes	Examples
Intrinsic motivation	Motivation intrinsic to use app	"I think the metrics on the whole were pretty good. And they, certainly from an eco-perspective, acceleration and braking, were probably pretty important aspects of that (08)".
	App scoring is fair	"When I was in the car with other people who aren't quite as safe drivers. The score was going down. So yeah, I think it was really accurate (26)".
	Fun	"You get coins, don't you? You can change your car colour or something (24)".
	App liked/ ease of use/ recommend	"I like the stars as well... the primary school thing, I think (03)".
	Should be built in car infotainment	"I was able to show that our trip where she would normally have criticized my braking, a 5 star rating for braking. So I sort of used it like that to say I'm not as bad as you think (08)".
		"Seeing my MPG go from about 36 to about 43 and thinking, oh, that's made a difference to me at a time when I could really do with it… kind of actually seeing that (11)".
		"My son is very paranoid about global warming sustainability. He's constantly on my case all the time, so if I could show him that I was committed to take that small step to help with that, to help with his future, not mine, but his future, then I think that would be something personal for me (17)".
		"I think I'd recommend it to somebody who felt like they could improve, like they wanted to improve their driving (12)".
		"You could do a deal with a manufacturer and actually have it as part of the main infotainment system… instead of having it on your phone (08)".
		"I did find it fun and I was telling my family about it (26)".
		"I guess it made me think about things I hadn't really thought about, but in a way that made it quite fun (12)".
Eco-driving	Greater awareness of eco-driving	"I didn't realize how harsh I accelerate until I started using the app. Cos I realize, I don't speed, but what I do is I get up to the speed limit quite fast instead. And I didn't realize that until I started using the app (21)".
	Specific improvement thanks to app	"I definitely watch like my rev counter more now (19)".
		"I didn't know it, I'm quite a hard accelerator, which I didn't realize (04)".
	Longevity	"The app has shown me that, yes, I brake quite, like, suddenly, or like, hard braking, I should say. So, then, that would say to me, yeah, probably I'm too close to the ones in front (23)".
		"I think my braking score went up once I was seeing the app and I was kind of more aware of that (26)".
		"I think it encouraged me to think like, oh, I am actually sticking to the speed limit and doing pretty well (12)".
		"I think it's something that is now, when I'm driving without the app on, conscious of it still (25)".
		"I wouldn't mind having it there. Would be quite interested to have a look at it every now and again (23).

continued

Themes	Codes	Examples
App positives	Leaderboard liked	"I thought the app itself was really easy to use… it was very straightforward and it was very appealing to look at (24)".
	Overall score good	"When we hadn't looked at the app for two weeks, it was quite nice to see where I'd been in that two weeks (17)".
	Tips useful	"I really liked it. Erm, so I could kind of, most of the events that showed, I kind of knew in advance where they would be (08)".
	Trips useful	"I thought it was really good, really accurate, and it was nice seeing the visual (26)".
		"I'm used to… running apps and stuff like that. That's kind of how you see runs laid out or whatever. So it felt like the most visually, like familiar in terms of the app (12)".
		"That conscious knowing that these are the people are on the road as well as you that are trying to improve their scores and become better drivers. I thought that was quite, it was a bigger picture sort of thing for me (15)".
		"I could see everyone else's scores I wanted to win and beat other people's scores. So yeah, I think it did encourage me slightly (14)".
		"I think the competition element was quite healthy (16)".
		"I was never at the top of the leaderboard, but that sort of spurred me on to make sure that I didn't drop any lower than that and then wondering who is this person, that number one that's always got 100% in just 100 for everything (27)".
		"I was able to move up in a positive way on the on the leaderboard (03)".
		"Yeah, it gives you a bit of a boost, I think when you see that you do well (12)".
		"So if you kind of advertised it as this, like competing with your friends or your family and see who is the better driver, I think I would be interested in that (13)".
		"If you could post your scoreboards straight to your socials or something like that might be a way of competing with friends (08)".
		"I really liked it, and me and my husband both had it, and then we had a bit of healthy competition going on (23)".
		"I guess it would encourage conversation with friends, maybe about how well you're doing on the leaderboard and stuff outside of using it (12)".
		"I think not knowing anybody and it was all like anonymized names anyway. I just kind of lost interest (19)".
		"Say if people have just passed their test for example like a bunch of 17-18 year olds and they could all get together and have it (04)".
		"I got tips and yeah, I did take them on board and I think it did increase sort of the scores (26)".
App niggles	Cornering not eco, but could be if affects wear tear	"It was wrong on the speed limits of a couple of roads where they've been redesigned near me, so it would forever say I was speeding and it was a journey I did every week and I was like, well, I know I'm not because I'm doing it and I'm checking my speeds and I'm very conscious and you're wrong, app! (11)".
	Emergency braking unfairness	"A couple of spots on the road there was two key points that kept saying I was speeding and I wasn't. So I'm not sure if it's the they haven't adjusted the speed limit (25)".
	Leaderboard disliked	
	Speeding inaccurate	"I think the speeding registered when you went 15 kilometres over the limit of the road that it recognised you are on (08)".

Themes	Codes	Examples
	Tips unhelpful	"It was just the same thing every time… and it was just constantly - Ohh try not to accelerate. Ohh try not to brake so hard (11)".
	App suggested improvements	"I don't find those very helpful just because I think the information is pretty self-explanatory (13)".
	Live feedback unwelcome	"They didn't tell you like how to improve it. They just said like, don't accelerate harshly and it's like OK, but how do I not if like I am doing it, how am I gonna fix that? (14)".
	Live feedback welcome	"I think that it sort of lacks a bit of a human element to it... it needs somebody who comes from some sort of driving organisation… that can give really practical tips, maybe a video of them like showing you how to brake well (12)".
		"Having short term challenges to, especially if there's a reward or something for it, then yeah, definitely (08)".
		"Different scores for each week with your friends so it could refresh the scores and then, say, if you had a bad week one week it wouldn't affect the scores next week (13)".
		"I think it's quite involved because there's so many different screens, so I feel like if there was just the one screen with the main driving score, the overall driving score, and then maybe the maps below or something. It felt like there was lots of different elements and I don't know that, on a daily basis, I would check that outside of the study (12)".
		"An alarm on the app to remind you to look at it the end of the day (12)".
		"I would have liked something there if something wasn't your fault (09)".
		"I think that would help me because like when I'm driving, obviously like when I'm driving with the app, I wasn't really thinking, oh, I wonder what I'm gonna get at the end of this day, I was kind of more focused on, like, how I was driving in that moment (14)".
		"Tells you if you're getting too close to the vehicle in front when you're on motorways and stuff, at high speeds and that, so that's the one I do keep on (05)".
		"But if you accelerating or, like, going over the speed limit, I think that's fair enough for it to ping, just for your own safety and others (06)".
		"I think like you're saying there, if there is a way to sort of notify you but not annoy you to death that you're not maybe doing that, then I think that's where you can probably get a bit of traction (10)".
		"I think I find a distraction. I just wanna swear at it (17)".
		"Something pinging would do my head in (18)".
		"To have something like that constantly telling me, would make me feel like I was doing a really bad job and it would just make me more nervous (04)".
		"I guess it must feel like having a back seat passenger, backseat driver kind of thing (16)".
		"I feel like it would, I think I'd be more tempted to sort of swipe something off my phone if it was coming up to say that I'd cornered badly or whatever (12)".

continued

Themes	Codes	Examples
Real world context	Privacy concerns	"Whilst I was trying to be aware of the app and not being marked down, I would be irritating the person behind me who was trying to get home faster than me (09)".
	Longevity plateau	"On some of the roundabouts, you have to get off pretty quick [i.e. accelerate rapidly]. And so I was like ohh, it's marked me down for that, but I'd have been there forever if I'd waited (25)".
	Miscategorisation fixes	"Can't always slow down or you trying to keep up with traffic a little bit, you know, not to annoy everybody else on the way home or way too work (10)".
	Motivation extrinsic examples	
	Peer pressure prevents eco driving	"You've got to put your foot down a little bit to actually stay within eco driving, otherwise you're sitting there for 15-20 minutes, wasting all that fuel just sitting there (05)".
		"As I started to lose points here and there, I sort of forgot about it a bit and didn't really care if it dropped down a bit more… if I need to be somewhere relatively quickly, then I won't take into account the sort of eco side of it (08)".
		"It's hit that point where I couldn't really improve much more. So irritating myself, I think. Really. Yeah. (09)".
		"It became quite a chore sort of looking at it every day because I come home, got to get the tea on and it's, it's another thing I've got to do (09)".
		"There was a little bit of a competition element, I think at first, to try and get it to a certain level. But then for me it kind of plateaued… I couldn't really improve it (10)".
		"It was really helpful using it for a few weeks… but then like for me that's probably enough for me to like become more eco without having the app downloaded anymore (19)".
		"You can't carry on with that kind of competitive rate for the rest of your life. You know how it is. It was good that that it all came to an end after three weeks, I think (03)".
		"Like the insurance thing I talked about where there's the lower premiums (11)".
		"For every 10,000 points you get will donate even if it's 10p, you know, to some kind of you know, sustainability or whatever (17)".
		"That would be even better, Yes. Definitely (17)".
		"That would be really good (15)".
		"Earn points which you can then use to put towards like an Amazon voucher for example (04)".
		"I think a discount on fuel would be quite effective, actually, especially in this whole fuel crisis thing (16)".
		"If it was advertising itself as it increases eco driving so reduce costs and it's better for the environment (19)".
		"Some sort of like car related stuff, So whether that's like, I don't know, like a Bluetooth connector (12)".
		"Would also be good to get smaller practical freebies that help with driving like petrol vouchers, windscreen covers or even just car de-icer (12)".
		"It might be nice if the 'overall score' had a 'you've saved ××× amount on petrol this week compared to the average driver' (12)".

Themes	Codes	Examples
Wider concerns	Secondary benefits of eco driving Gamification distorts More than eco-driving (bigger picture) Thoughts about study organisation	"How fast could you go without them notifying the authorities? Would they never notify the authorities or anything like that? (08)". "I know personally if, if I saw an offer to have a black box I would 100% not choose that (16)". "If people wanted to treat it like a game, so they have to find some way to manipulate the data to make them seem like they're driving better than they actually are. I don't know whether the scores is like an average on the miles you were doing for each journey. For example, if you start on the motorway for 300 miles doing 70 miles an hour. But if you're going to get a better driving score, but you just burning fuel for just the sake of it, which is not very fuel, you know, eco friendly (16)". "If you were really that interested in scoring high, then if you drove like a hooligan on a trip, you just say you're a passenger and it would take you out of the score (08)". "We've recently moved house and positioned ourselves so that I can walk to work. I can walk my kids to nursery, I can walk them to school, so that's kind of the way we do it (11)". "If you were driving, which… shouldn't be driving 50 in a 30, it would actually probably be more eco but it wouldn't be safe (08)". "There was one particular person who only seemed to have travelled 40 kilometres… other people seemed to be doing thousands, so it was, it is quite difficult to know (03)". "A lot of other people join, then that affects your position on the leaderboard, doesn't it? (09)". "Saw that there was people like on 100% and they only done like 14 miles, and then there's others that had done 2-3 or thousands of miles as well. And it it's like, well, it's, it's not really a scoreboard. So it, it lost all its legitimacy (05)". "It was either that or mow down the pedestrian in front of me (11)". "I had passengers in the car and I had to brake. Obviously it was one of those moments where you have no choice. Like we said before. And I just thought in my head. Oh God, the app. And it just came up on the app later that night and I thought ohh look what I've done. (Laughs) (15)". "Driving home in rush hour and you're having to slam your brakes on because somebody's done something stupid (09)".
It's just not for me	App didn't work for me	"For me personally, I don't think I really changed how I drove throughout it (24)". "I was very much more aware of what I was doing when I knew that I could see it on the app, erm, but it didn't really make me alter my ways (04)". "It just got annoying that it was marking me down on other people pulling out and things like that (05)". "It (the score) just got lower and lower and I found myself getting more frustrated that it was affecting my driving making it worse (05)". "Personally, I just, I'd rather not have to spend like more time my phone and I already have (19)". "I want to enjoy… driving without scrutinising myself all the time (09)". "I wasn't able just to relax into, at my normal driving pattern (05)". "It took away, it made it harder for me to drive (05)".

Towards a student-centered education: Validation of the Italian version of the Conceptions of Learning and Teaching questionnaire

Giulia Rampoldi[1], Laura Montelisciani[1], Johanna C.G. Jacobs[2], Selena Russo[1], Stefano Ardenghi[1, 3],
Marco Bani[1], Laura Antolini[1], Fabrizio Consorti[4], Maria Grazia Strepparava[1, 3]

[1] School of Medicine and Surgery, University of Milano-Bicocca, Monza (MB), Italy
[2] Utrecht Center for Research and Development of Health Professions Education,
University Medical Center Utrecht, Utrecht, The Netherlands
[3] Fondazione IRCCS San Gerardo dei Tintori, Monza (MB), Italy
[4] Department of Surgery, Sapienza University of Rome, Italy

mariagrazia.strepparava@unimib.it

❧ ABSTRACT. Il presente studio osservazionale e multicentrico ha lo scopo di tradurre e validare la versione italiana del questionario *Conceptions of Learning and Teaching* (*COLT-IT*). Lo studio ha coinvolto 394 docenti afferenti a 13 diverse Scuole mediche italiane. Le analisi statistiche eseguite, rispetto alla versione originale dello strumento, hanno portato all'eliminazione di uno dei tre fattori e di 11 dei 18 item. Il COLT-IT risulta quindi essere composto da 7 item e due fattori che hanno confermato i nomi originali di Teacher centredness (TC) e Appreciation of active learning (AL). Il COLT-IT è uno strumento valido, affidabile e di facile somministrazione, utile a sostenere nel panorama della medical education internazionale la transizione da un modello formativo tradizionale centrato sul docente ad un modello attivo centrato sullo studente.

❧ SUMMARY. *Many medical schools have transitioned from traditional teacher-centred education to active student-centred education. The Conception of Learning and Teaching (COLT) questionnaire investigates teachers' conceptions of learning and teaching in student-centred medical education. This observational and multicentred study aims to validate the Italian version of the Conceptions of Learning and Teaching (COLT-IT) questionnaire and assess the Italian medical educators' learning and teaching conceptions. To develop the COLT-IT, a back-translation was performed. The COLT-IT and socio-demographic questionnaires were electronically distributed to educators across 13 Italian medical schools. Analyses included explorative factor analysis (EFA), confirmatory factor analysis (CFA), and reliability analysis. A total of 394 medical teachers completed the survey. Although the EFA suggested retaining three factors, statistical conditions led to the exclusion of eleven of the original eighteen items and one factor. The CFA confirmed the bi-dimensional structure of the COLT-IT. The two-factor scale retained the original naming of subscales: Teacher centredness (TC) and Appreciation of active learning (AL). Approximately 50% of participants exhibited high levels of AL, while scores in the TC subscale were more varied. No differences emerged based on gender, age, or the geographical location of the medical school. The 7-item COLT-IT is a reliable, valid, robust, and easy-to-administer tool for promoting and monitoring the implementation of a student-centred approach in medical education.*

Keywords: *Cooperative learning, Higher education, Teacher attitudes*

INTRODUCTION

In contemporary society medical students are requested to develop sound technical and clinical medical knowledge as well as effective communication and relational skills (Dent, Harden & Hunt, 2017). In response to these evolving demands, many medical schools have transitioned from the traditional teacher-centred education (TCE) to student-centred education (SCE) (McLean & Gibbs 2010). TCE places teachers and lecturers at the centre of the learning process, with students functioning as passive knowledge repositories (Harden, Sowden & Dunn, 1984). On the other hand, SCE is characterized by a greater attribution of responsibility to students, who become active elements of their own learning processes, while teachers take on the role of facilitators of these processes (Attard, Di Iorio, Geven & Santa, 2010). This paradigm shift has led to the introduction of innovative teaching methods in many medical curricula, including problem- and case-based learning (McLean, 2016; Nundy, Kakar & Bhutta, 2022), team-based learning (Michaelsen, Sweet & Parmelee, 2011), communication skills training (Rotthoff et al., 2011), and medical humanities (Wald, McFarland & Markovina, 2019).

Research indicates that SCE is associated with numerous advantages and positive outcomes for students and teachers in healthcare (Hopper & Brake, 2018). Compared to TCL, SCE has proven to enhance academic motivation, satisfaction, and confidence and fostering deeper learning in students (Covill & Cook, 2019; Grijpma, Mak-van der Vossen, Kusurkar, Meeter & de la Croix, 2021). Furthermore, it has been linked to improved team working abilities, effective communication (Li, Wang, Zhu, Zhu & Sun, 2019; Sulaiman, Shahimi & Zakaria, 2021), clinical reasoning (Ulfa, Igarashi, Takahata, Shishido & Horiuchi, 2021), critical thinking (Xhomara, 2022), patient-centredness (Bombeke et al., 2010) and greater problem-solving performances (Kim, Song, Lindquist & Kang, 2016). Despite these promising benefits, SCE is not widely implemented at the university level yet (McLean & Gibbs, 2010). A critical issue associated with the implementation of the SCE is represented by teachers' explicit and implicit conceptions of learning and teaching (Postareff, Lindblom-Ylänne & Nevgi, 2008). Learning and teaching conceptions encapsulate teachers' attitudes and beliefs regarding the learning and teaching processes, though these may not be directly manifested in their teaching strategies, which

represent teachers' observable didactic behaviours (Pajares, 1992). Nevertheless, changes in teaching strategies can only occur when shifts in teachers' conceptions and attitudes toward education take place (Trigwell & Prosser, 1996). Despite the evolving landscape of education, many lectures and academics remain attached to traditional curriculum, making it challenging to persuade them to adopt innovative teaching methods (Kember, 2009). Therefore, assessing teachers' conceptions becomes imperative to design interventions and strategies aimed at promoting SCE.

Several instruments assessing educators' teaching and learning conceptions are available, including: 1) the *Learning Inventory* (Bolhuis & Voeten 2004); 2) the *Approaches to Teaching Inventory* (Trigwell, Prosser & Waterhouse, 1999); 3) the Motiva*tion Strategies for Learning Questionnaire* (Postareff, 2007); 4) the *Inventory of Teaching Patterns* (Donche, De Maeyer & Van Petegem, 2007); 4) the *Teaching Perspectives Inventory* (Pratt & Collins, 2001) and 5) the *Conceptions of Learning and Teaching* (Jacobs et al., 2012). Among these tools, the *Conceptions of Learning and Teaching* (*COLT*; Jacobs et al. 2012) has been specifically developed to investigate teachers' conceptions of learning and teaching in the context of the student-centred medical education. Notably, there are no other validated instruments measuring teachers' conceptions of learning and teaching available in Italian. The availability of the Italian version of COLT (COLT-IT) represents a promising instrument that can provide valuable insights for promoting and monitoring the implementation of more student-centred teaching in the Italian educational context. This could help bridge the gap with the international context and establish a standardised Italian pathway for medical training (Carvalho, Dane & Whicker, 2021; Jacobs et al., 2014).

AIMS

This study aimed to achieve three primary objectives: 1) translate and validate the Italian version of the COLT (Jacobs et al. 2012); 2) verify its psychometric properties; 3) assess conceptions of learning and teaching in a diverse sample including lecturers, professors, laboratory tutors and clinical mentors (healthcare professionals who work in hospital and clinical structures) from various Italian medical schools.

In validating the Italian version of COLT, our expectations included confirming the three-factor structure of the tool

and finding no significant differences in the data across medical schools in different Italian regions. Additionally, we anticipated observing a higher inclination towards teacher-centeredness in our sample compared to an appreciation for active learning.

MATERIALS AND METHOD

Materials

Self-report questionnaires collected participants' socio-demographic and professional information (gender, age, university, academic role, seniority) and their conceptions of learning and teaching in medical education. The latter was measured with the Italian translation of the COLT (Jacobs et al. 2012). The original COLT was built using the Delphi method starting from the multifactorial structure of the *Learning Inventory* (Bolhuis & Voeten, 2004) and a confirmatory factor analysis (CFA) was performed to analyse the data. It consists of 18 items evaluated on a 5-point Likert scale (1 = completely disagree; 5 = completely agree). The questionnaire presents three subscales: 1) the Teacher centeredness (TC) scale (8 items, Cronbach's α = .73) evaluates the teacher's orientation in considering teaching as a transmission of knowledge with student in a passive position; 2) the Appreciation of active learning (AL) scale (5 items, Cronbach's α = .57) evaluates the appreciation of the constructivist vision of learning based on a conceptual change of students and 3) the Orientation to professional practice (P) scale (5 items, Cronbach's α = .63) evaluates teachers' conceptions regarding the integration of future professional practice during the years of undergraduate medical education. In order to develop the Italian translation and adaptation of the COLT, a back-translation process was performed (Brislin, 1986) using four bilingual translators.

Method

The present research is an observational, multicentred and non-randomized study. Participants were recruited electronically through a convenient snowballing non-probabilistic sampling method. The Italian Society of Medical Education (SIPeM) and the Permanent Italian

Conference of the Directors of Undergraduate Medical Schools invited 13 Medical Degree Program Directors via institutional emails to share the survey link among their lectures and professors.

The invited medical schools were selected to ensure equal distribution between the Northern and the Southern regions of the country.

Ethical approval

The study was approved by the Ethics Committee of the University of Milano - Bicocca (Protocol number 0109004/19 of 12/12/2019).

Strategy of data analyses

The dataset was preliminarily tested performing an analysis of multivariate outliers using the Mahalanobis' distance (set at $p<.001$) (Leys, Klein, Dominicy & Ley, 2018). The data analyses were based on the standard procedure for instrument development (Matsunaga, 2010). An exploratory factor analysis (EFA) was initially performed using the factor analysis method (FA) to extract the factors followed by orthogonal rotation of factors using Varimax rotation. As questionnaire items are ordinal variables and given that the FA model assumes that manifest variables are linear combinations of continuous common factors, the weighted least squares method (WLS) (Muthén, 1984) for factor analysing ordinal variables was used. This method assumes an unobservable normally distributed continuous variable underlies each observed ordinal variable in the population. The number of factors to be retained was guided by parallel analysis (PA) (Horn, 1965). Only the items that met statistical conditions of primary (less than |.45|; Comrey & Lee, 1992) and secondary loadings (greater than |.20|) were retained. The output obtained from the EFA was used as a baseline to perform subsequent psychometric analyses.

Then, a confirmatory factor analysis (CFA) was performed (Gagne & Hancock, 2006) using the diagonally weighted least squares (DWLS) estimation specifically designed for ordinal data. To evaluate the goodness of fit of the model tested in the CFA, the following indexes were calculated: χ^2 ($p>.05$), Root Mean Square Error of Approximation (RMSEA) (acceptable<.10, good<.08, very good<.05), Standardized Root

Mean Square Residual (SRMR<.05), Comparative Fit Index (CFI>.90) and Tucker-Lewis Index (TLI>.90) (Morin, Marsh & Nagengast, 2013). The assumption of uni-dimensionality of the model (M1) was initially tested (Judd, Jessor & Donovan, 1986). This analysis was performed to evaluate the discriminant validity and to compare the goodness-of-fit indices of a single-factor measurement model with a model consisting of all the instrument's characteristics (Judd et al., 1986). Following the indications provided by the EFA, two other CFA models were performed to achieve good model adaptability indexes (M2 and M3).

The psychometric properties of the COLT-IT were then analysed. To acquire a quantitative representation of the Italian medical teachers' student-centeredness or teacher-centeredness, the χ^2 analysis was performed to evaluate the possible differences in the answer distributions for geographical location of the medical school (North and South), gender (female and male), and age (less equal 55 and over 55).

RESULTS

Participants

A total of 401 medical teachers completed the survey, out of which seven multivariate outliers were identified and subsequently excluded. The final sample comprised 394 participants, which is considered a good sample size for validation procedures, where 21.9 participants per item are recommended (Suresh & Chandrashekara, 2012). The participants aged between 27 and 73 years ($M = 55.67$; $SD = 9.31$) and their seniority extended from a minimum of 1 year to a maximum of 45 ($M = 22.18$, $SD = 10.45$). Table 1 reports the socio-demographic characteristics of the participants.

Exploratory factor analyses (EFA)

An EFA was conducted from the data collected from the 394 participants. We checked that each item had at least one polychoric correlation greater than |.30|. The Kaiser-Meyer-Olkin (KMO) test was .82, indicating that the sample was adequate. The Bartlett's test results statistically significant ($\chi^2 = 3106.735$; $p = .000$), indicating that the relationship among the variables was strong and the data were suitable to conduct an EFA.

The PA suggested to consider three factors (Cattell, 1966). An accepted rule of thumb is to consider 5 to 10 cases per parameter (Kline, 2011). Accordingly, three-factor model in this study required a minimum sample size of 45 or 90. In this analysis, to obtain a simpler solution with an easier interpretation of its result, a Varimax rotation was performed. Two criteria were used to select items to be retained: each item must not have the primary loading less than |.45| (Comrey & Lee, 1992) and the high secondary loading greater than |.20|. Only the items that met the two conditions simultaneously were included in the analysis. After the items inclusion/exclusion process a two-factor model emerged. The two factors were named as follows: Factor 1, Teacher centeredness (TC) included four items (items 1-2-4-5) and showed factor loadings ranging from .54 to .61; Factor 2, Appreciation of active learning (AL) included three items (items 10-13-18) with a factor loading from .61 to .89 (see Table 2).

As for the variance explained by the two factors, 26% of the variance was explained by TC and 16% by AL, totalling an explained variance of 42%.

Confirmatory factor analyses (CFA)

The hypothesis of uni-dimensionality of the model (M1) was preliminarily tested. The analyses of goodness of fit indexes revealed a general poor fit between the model and empirical data [$\chi^2_{(135)} = 1502.10$, $p = .000$, TLI = .735, CFI = .766, RMSEA = .165, $p = .000$, SRMR = .151] and suggested to reject the hypothesis of a single latent factor. Then, the three-dimensional model (M2) was tested as indicated by the questionnaire original version (Jacobs et al., 2012). Results of CFA reveal that also M2 is characterized by numerous goodness-of-fit indices far from the acceptability threshold [$\chi^2_{(132)} = 747.49$, $p = .000$, TLI =. 878, CFI = .895, RMSEA = .112, $p = .000$, SRMR = .112] (see Figure 1). Considering the EFA results, a two-dimensional model (M3) was tested. The factor 1 and 2 have a low correlation (.04). Results of CFA suggested the acceptability of the factorial model underlying the two-dimensional model M3: $\chi^2_{(13)} = 17.56$, $p = .174$, TLI = .991 CFI = .994, RMSEA = .031, $p = .804$, SRMR = .045. In particular, the saturation values of all items were medium-high ranging from $\lambda = .51$ to $\lambda = .88$ (see Figure 2). The theoretical interpretation of the COLT-IT factors supported

Table 1 – Socio-demographics characteristics of the participants

	Total sample (N = 394)	
	N	*%*
Gender		
Female	170	43
Male	223	57
Academic role		
Full professor	115	29.2
Associate professor	166	42.1
Researcher	103	26.1
Research fellow	3	.8
Lecturers	7	1.8
University geographical location		
North Italy (*N* = 9)	199	50.6
South Italy (*N* = 4)	194	49.4

the maintenance of the original naming: Teacher centredness (TC) and Appreciation of active learning (AL).

Reliability, convergent validity, and descriptive statistics of COLT-IT

The convergent validity of the measurement model has been assessed by the average variance extracted (AVE) and the composite reliability (CR). The discriminant validity was assessed by the heterotrait-monotrait ratio (HTMT) of the correlations (Henseler, Ringle & Sarstedt, 2015). CR values between .7 and .9 are considered satisfactory (Nunnally, 1978). In this analysis it was .84. For constructs with categorical items, ordinal coefficients alpha was calculated (Zumbo, Gadermann & Zeisser, 2007). They were .67 for TC and .78

for AL. The AVE is calculated from polychoric correlation matrix. It was .34 for TC factor and .56 for AL factor. AVE values greater than .50 indicate that, on average, the construct explains more than half of the variance of its indicators. AVE values less than .50 indicate that the error in the items is greater than the variance explained by the construct. In order to clearly discriminate between two factors, the HTMT should be significantly smaller than 1. If the value of the HTMT is higher than 1, there is no discriminant validity. In this study HTMT between TC factor and AL factor was .14 showing a good discriminant validity.

Descriptive statistics are reported in Table 3. No significant differences in the answer distributions of the two COLT factors emerged for geographical location of the medical school, gender, and age, supporting the empirical adoption of the COLT-IT.

Table 2 – Results of EFA: two dimensions composed of 3 and 4 items respectively were accepted

Items	F1	F2	F3	Communalities	Primary>|.45|	Secondary<|.20|
COLT_1	**.54**	.04	.13	.31	.54	.13
COLT_2	**.59**	−.02	−.10	.36	.59	−.10
COLT_3	.70	.31	−.07	.59	.70	.31*
COLT_4	**.61**	.00	.05	.37	.61	.05
COLT_5	**.58**	.15	.18	.39	.58	.18
COLT_6	.49	.09	.33	.36	.49	.33*
COLT_7	.62	−.21	.16	.45	.62	−.21*
COLT_8	.37	−.34	.02	.25	.37*	−.34*
COLT_9	−.29	.58	−.02	.42	.58	−.29*
COLT_10	.08	**.61**	.18	.41	.61	.18
COLT_11	.13	.59	.27	.44	.59	.27*
COLT_12	.08	.55	.22	.36	.55	.22*
COLT_13	−.03	**.66**	.18	.47	.66	.18
COLT_14	.08	.26	.60	.43	.60	.26*
COLT_15	.11	.42	.81	.84	.81	.42*
COLT_16	.16	.55	.52	.60	.55	.52*
COLT_17	.28	.64	.27	.56	.64	.28*
COLT_18	−.04	**.89**	.14	.81	.89	.14

Note. * the item not meet the condition and was excluded.

DISCUSSION

This psychometric validation study aimed at validating the Italian version of the *Conceptions of Learning and Teaching* (*COLT-IT*) instrument and at exploring its dimensions in a group of Italian medical teachers. The findings support the use of the 7-item COLT-IT, which comprises two dimensions: Teacher-centredness (TC) (4 items) and Appreciation of active learning (AL) (3 items).

The results of this study revealed a reliable, valid, and solid questionnaire, which loaded onto two factors rather than the three identified in the original validation work (Jacobs et al., 2012). Although the results of PA initially suggested retaining three factors, the statistical criteria for retaining items led to the exclusion of eleven items (original items 3, 6, 7, 8, 9, 11, 12, 14, 15, 16, 17) and subsequently the elimination of one factor. The CFA confirmed the bi-dimensional structure of the questionnaire. Notably, a bi-dimensional structure has also

Figure 1 – CFA results of the original three-dimensional model of measurement of COLT (M2)
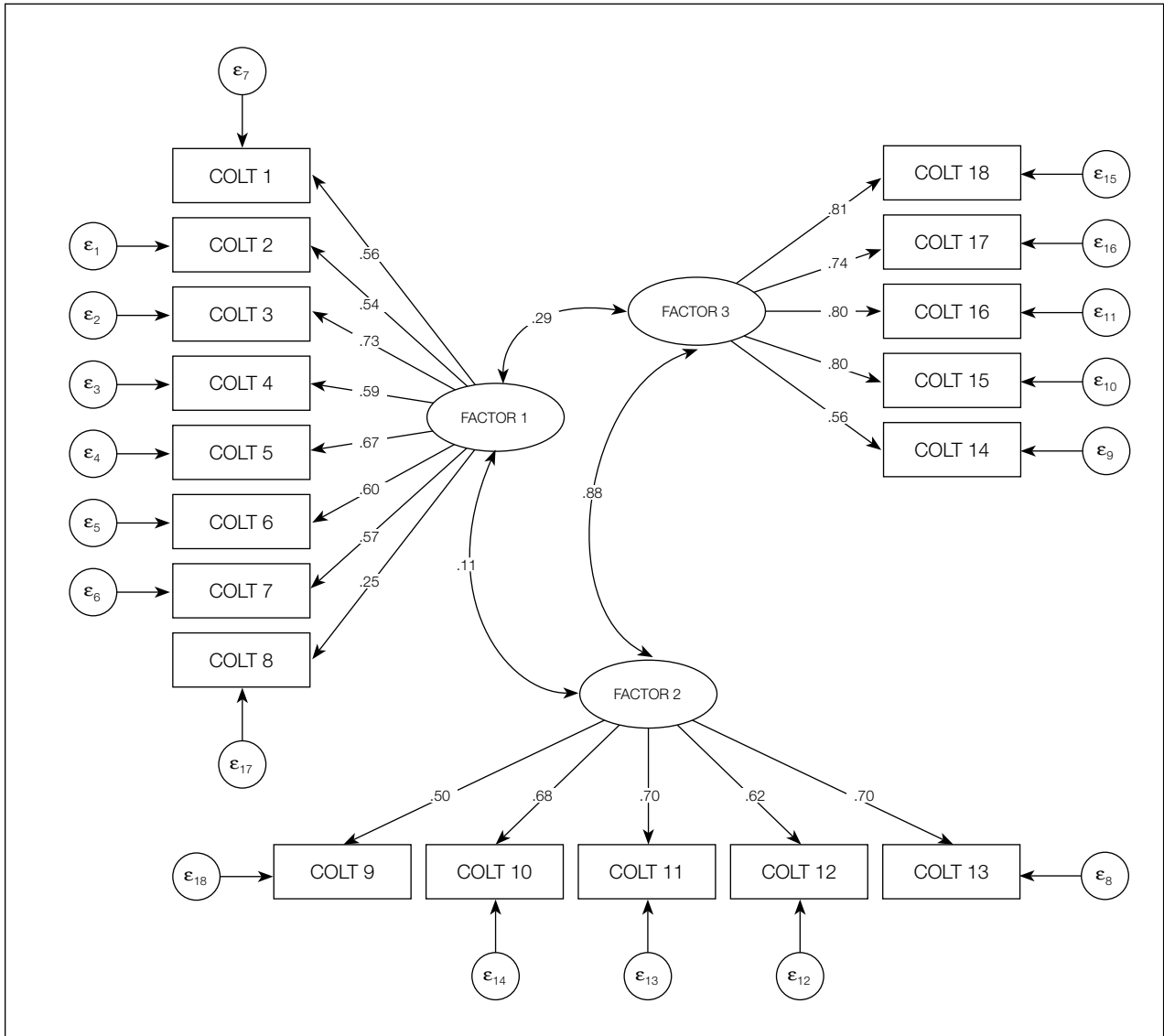

Figure 2 – CFA results of final two-dimensional model of measurement of COLT (M3)
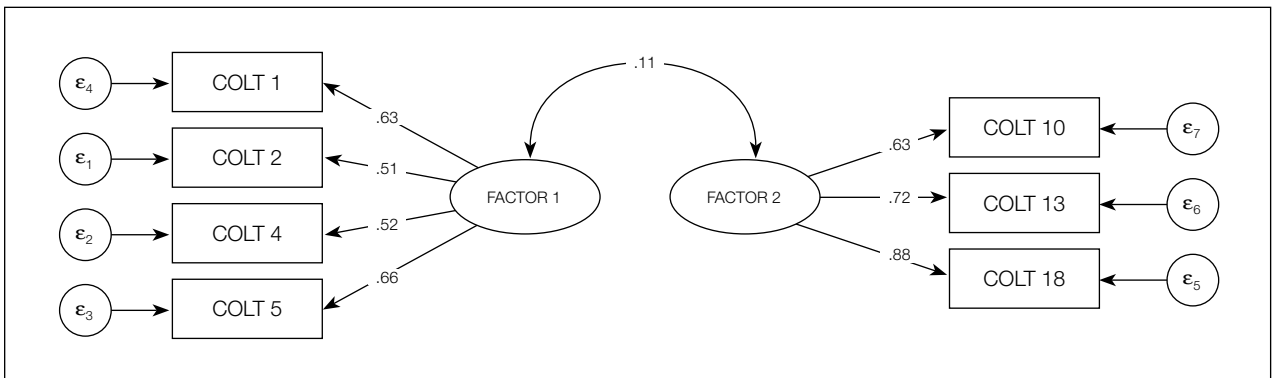
Table 3 – Demographic characteristics and psychometric proprieties of the COLT-IT subscale: Teacher centredness (TC) and Appreciation of active learning (AL)

Variables	Teacher centerdness (TC)												Appreciation of active learning (AL)											
	1		2		3		4		5		p-value		1		2		3		4		5		p-value	
	n	%	n	%	n	%	n	%	n	%			n	%	n	%	n	%	n	%	n	%		
Area											.070												.238	
North	6	3.2	20	10.8	42	22.6	62	33.3	56	30.1			0	.0	5	2.7	20	10.8	73	39.2	88	47.3		
South	6	3.2	20	10.8	33	17.7	45	24.2	82	44.1			3	1.6	8	4.3	27	14.5	63	33.9	85	45.7		
Gender											.887												.929	
Males	4	2.5	15	9.4	33	20.8	49	30.8	58	36.5			1	.6	7	4.4	21	13.2	56	35.2	74	46.5		
Females	7	3.3	25	11.7	42	19.7	58	27.2	81	38.0			2	.9	6	2.8	27	12.7	79	37.1	99	46.5		
Age											.972												.373	
≤55	5	3.4	18	12.2	29	19.6	43	29.1	53	35.8			0	.0	8	5.4	20	13.5	54	36.5	66	44.6		
>55	7	3.2	22	10.0	45	20.5	63	28.6	83	37.7			2	.9	5	2.3	26	11.8	81	36.8	106	48.2		

been reported in the COLT version for postgraduate students (Pacifico et al., 2021). The theoretical interpretation of the two-factor scale supported retaining the original naming: Teacher centredness TC and Appreciation of active learning AL. Despite the exclusion of 11 items and the disparities observed with both the original COLT and the validation by Pacifico and colleagues (2021), it is noteworthy that the two factors of the COLT-IT exhibit high consistency, as evidenced by the Cronbach's alpha indices, which are higher than those reported in previous validations.

The absence of the original factor in the COLT-IT is notable. This factor represents the teacher's belief regarding the integration of future professional practice during undergraduate medical education. Over the past few decades, Italian medical programs have undergone significant innovation and reform aimed at professionalising teaching (Consorti, 2018). However, despite these efforts, most Italian medical schools have remained predominantly theoretical-oriented compared to their counterparts in other European countries (Snelgrove et al., 2009). Hence, the absence of the P factor in the COLT-IT appears to align with this cultural tendency. The P factor items appear to be misunderstood by Italian medical educators, who may not fully-grasp the concept of learning to serve practice. Instead, Italian medical teachers tend to prioritise the professionalization of medical education by implementing innovative active teaching methods centred on students, aimed at enhancing students' soft skills such as problem-solving, team-working, and critical thinking (Familiari, Violani, Relucenti & Heyn, 2013). It is possible that Appreciation of active learning (AL) among Italian teachers conceptually overlaps with their Orientation to professional practice (P). This hypothesis is supported by the high correlation between these two factors observed in the second model tested in the CFA. Moreover, similar overlapping between the AL and P has been found in the COLT version for postgraduate students, where the authors combined the P and AL factors in the A-P factor (Pacifico et al., 2021). In Pacifico's study, out of the 5 items originally in the P factor, three were removed (items 15, 17 and 18), and the remaining two (items 14 and 16) loaded on the combined factor A-P. Interestingly, in our study, items 15 and 17 were also removed. However, in contrast to Pacifico's findings, items 14 and 16 were removed in our study and did not shift to the AL factor, while item 18 did shift to AL. This discrepancy between our results and those of Pacifico may be attributed to differences in the target population. Specifically,

item 18 ("Discussing topics with each other helps students to learn how to deal with different points of view, so as to gain a deeper understanding") was removed in Pacifico's study due to its significant disturbance. The author suggested that for residents, "the emphasis at this point is on actual activities that induce learning and not on theoretical discussion" (Pacifico et al., 2021).

It is however, possible that the absence of the P factor and the exclusion of eleven items in our study and the overlapping of the factor P and factor AL in Pacifico's study could also be attributed to the EFA used in this study, which differed from the original questionnaire development process where a Delphi method employed (Jacobs et al., 2012). Further studies conducted in different cultural and learning contexts are necessary to validate this hypothesis. As for the 5 items originally belonging to the AL factor, three items have been removed in our analysis (item 9, 11 and 12). In Pacifico and colleagues (2021), only item 9 was removed. This difference in results with Pacifico's one may be linked to the different study population. Item 11 ("Small group learning motivates students to study") and item 12 ("I think it is more important for students to be able to analyse and critically appraise subject matter than to memorise facts") describe educational activities that in the Italian context are traditionally seen as both more appropriated and feasible for residents rather than undergraduate students (Consorti, 2018).

When comparing the eight items of the original TC factor, four items were removed in our study (items 3, 6, 7, and 8), whereas in Pacifico's study, only one item was removed (item 8). One possible explanation for this difference is once again related to variations in the study populations.

Items 3 ("Students learn best when the learning process is guided by an expert who has an overview of the field of interest"), 6 ("As a teacher, I have to indicate clearly what is important and what is less important for the students to know"), and 7 ("I think that as an expert in my field, I am eminently suitable to transmit my knowledge to students and that students should not have to look up that knowledge for themselves") express strong positions regarding learning and teaching, which may be traditionally associated with practical and clinical competencies rather than theoretical knowledge.

Analysing the distribution values of the COLT-IT dimensions, it was observed that nearly half of the Italian medical teachers scored high on the AL factor, with no significant differences based on gender, age, or medical

school geographical location. Conversely, the distribution of agreement with the TC factor was more varied, possibly reflecting the ongoing educational paradigm shift characterising Italian medical schools. No difference emerged for gender, age, or medical school geographical location. Interestingly, the absence of gender and age differences diverge from existing data indicating a tendency for female teachers to prefer active teaching (Norton, Richardson, Hartley, Newstead & Mayes, 2005), while male teachers tend to favour teacher-centredness (Jacobs et al., 2015). Previous studies have also shown that more teaching experience, often associated with age and academic status, is linked to a more student-centred approach (Jacobs et al., 2015; Sadler, 2012). Furthermore, teachers' conceptions of learning and teaching have been proven to be influenced by several contextual (e.g. type of medical school, curriculum, department management, leadership style) and personal factors (e.g. agency, work engagement, motivation, content expertise (Jacobs et al., 2014, 2015, 2016, 2020). As the present study did not explore these factors, further investigations are warranted to analyse the personal, professional, and cultural variables associated with Italian medical teachers' conceptions of learning and teaching.

Practice implications

The availability of this concise questionnaire can cater to both research and education assessments. The COLT-IT enables the measurement, systematization, and dissemination of didactic innovation efforts within Italian medical schools. Furthermore, the COLT-IT can serve as a valuable tool for promoting and monitoring the educational paradigm shift to a more SCE, informing faculty development strategies. It also provides an opportunity to bridge reduce the gap with the international context and establish a uniform Italian path to medical training (Familiari & Consorti, 2013).

Strenghts, limitations and future directions

The present study has several limitations that warrant acknowledgment. Firstly, although we have evaluated the convergent validity using the average variance extracted, it would be appropriated to evaluate it with other correlated

constructs including patient-centredness, self-efficacy, and burnout. Secondly, this study relied solely on self-report measures. Utilising an ecological study design characterised by external observation of teachers' didactic practices could provide additional insights into the validity of the COLT-IT. Additionally, the absence of other validation studies of the COLT restricts both the statistical and cultural discussion surrounding its implementation and interpretation.

Furthermore, we opted to conduct exploratory factor analysis (EFA) and confirmatory factor analysis (CFA) on the same sample instead of dividing the population into two groups. This decision was made to avoid a significant reduction in the sample size. Dividing the sample into two groups would have resulted in too few subjects per item. A similar approach was taken by Pacifico and colleagues (2021), who had a comparable sample size to ours. These limitations should be considered when interpreting the findings of this study.

While recent research (Jacobs et al., 2020) has supported the validity of using COLT internationally, this study represents the first validation of the COLT in a different language. A notable strength of this study is the demonstration that the COLT-IT is a concise questionnaire with optimal statistical characteristics. The availability of such a questionnaire, consisting of only 7 items and easy to administer, has the potential to encourage increased research in this area.

Future qualitative and quantitative studies are recommended to delve deeper into the external validity of the COLT-IT and to explore teachers and students' opinions regarding the implementation of a student-centred approach in Italian medical schools. Additionally, the COLT-IT could be used to investigate variables associated with teaching attitudes, enabling the implementation of targeted faculty development and training strategies. In a broader context, further translation and validation studies would facilitate the assessment of cultural factors' impact on teachers' conceptions of teaching and learning.

CONCLUSION

The Italian 7-item version of the COLT emerges as a valid, reliable, and sensitive instrument for evaluating teachers' conceptions of learning and teaching. Italian medical schools stand to benefit from its use in monitoring and promoting the educational paradigm shift towards a more

SCE, thereby informing strategies for faculty development. Notably, this study represents the first validation of the COLT in a different language, and its availability may catalyse expanded international research on teachers' conception of learning and teaching, as well as on faculty development strategies.

References

ATTARD, A., DI IORIO, E., GEVEN, K., & SANTA, R. (2010). Student-Centred Learning: Toolkit for students, staff and higher education institutions. *European Students' Union (NJ1)*.

BOLHUIS, S., & VOETEN, M.J. (2004). Teachers' conceptions of student learning and own learning. *Teachers and Teaching*, *10* (1), 77-98.

BOMBEKE, K., SYMONS, L., DEBAENE, L., DE WINTER, B., SCHOL, S., & VAN ROYEN, P. (2010). Help, I'm losing patient-centredness! Experiences of medical students and their teachers. *Medical Education*, *44* (7), 662-673.

BRISLIN, R.W. (1986). *The wording and translation of research instruments*.

CARVALHO, H., DANE, F.C., & WHICKER, S.A. (2021). Conceptions of learning and teaching for faculty who teach basic science. *Medical Science Educator*, *31* (2), 745-751.

CATTELL, R.B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research*, *1* (2), 245-276.

COMREY, A.L., & LEE, H.B. (1992). *A first course in factor analysis*. New York: Psychology Press.

CONSORTI, F. (2018). *Didattica professionalizzante nei corsi di laurea in medicina*.

COVILL, L., & COOK, J. (2019). Comparison of academic performance in traditional and flipped classrooms and students' attitudes of the flipped experience. *Journal of Allied Health*, *48* (1), 1E-7E.

DENT, J., HARDEN, R.M., & HUNT, D. (2017). *A practical guide for medical teachers*. Elsevier Health Sciences.

DONCHE, V., DE MAEYER, S., & VAN PETEGEM, P. (2007). Teachers' conceptions of learning and teaching and their effect on student learning. *Paper presented at the British Educational Research Association Annual Conference*, 5, 8.

FAMILIARI, G., & CONSORTI, F. (2013). The best evidence medical education and the essential skills on medical teaching: Important keys for medical education internationalization. *Med Chir*, *59*, 2662-2663.

FAMILIARI, G., VIOLANI, C., RELUCENTI, M., & HEYN, R. (2013). La realtà internazionale della formazione medica. *Medic*, *21* (1), 53-59.

GAGNE, P., & HANCOCK, G.R. (2006). Measurement model quality, sample size, and solution propriety in confirmatory factor models. *Multivariate Behavioral Research*, *41* (1), 65-83. doi.org/10.1207/s15327906mbr4101_5

GRIJPMA, J.W., MAK-VAN DER VOSSEN, M., KUSURKAR, R.A., MEETER, M., & DE LA CROIX, A. (2022). Medical student engagement in small-group active learning: A stimulated recall study. *Medical Education*, *56* (4), 432-443.

HARDEN, R.M., SOWDEN, S., & DUNN, W.R. (1984). Educational strategies in curriculum development: The SPICES model. *Medical Education*, *18* (4), 284-297.

HENSELER, J., RINGLE, C.M., & SARSTEDT, M. (2015). A new criterion for assessing discriminant validity in variance-based structural equation modeling. *Journal of the Academy of Marketing Science*, *43* (1), 115-135.

HOPPER, M.K., & BRAKE, D.A. (2018). Student engagement and higher order skill proficiency: A comparison of traditional didactic and renewed integrated active learning curricula. *Advances in Physiology Education*, *42* (4), 685-692.

HORN, J.L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, *30* (2), 179-185.

JACOBS, J.C., MUIJTJENS, A.M., VAN LUIJK, S.J., VAN DER VLEUTEN, C.P., CROISET, G., & SCHEELE, F. (2015). Impact of institute and person variables on teachers' conceptions of learning and teaching. *Medical Teacher*, *37* (8), 738-746.

JACOBS, J.C., VAN LUIJK, S.J., GALINDO-GARRE, F., MUIJTJENS, A.M., VAN DER VLEUTEN, C.P., CROISET, G., & SCHEELE, F. (2014). Five teacher profiles in student-centred curricula based on their conceptions of learning and teaching. *BMC Medical Education*, *14* (1), 220.

JACOBS, J.C., VAN LUIJK, S.J., VAN BERKEL, H., VAN DER VLEUTEN, C.P., CROISET, G., & SCHEELE, F. (2012). Development of an instrument (the COLT) to measure conceptions on learning and teaching of teachers, in student-centred medical education. *Medical Teacher*, *34* (7), e483-e491.

JACOBS, J.C., VAN LUIJK, S.J., VAN DER VLEUTEN, C.P., KUSURKAR, R.A., CROISET, G., & SCHEELE, F. (2016). Teachers' conceptions of learning and teaching in student-centred medical curricula: The impact of context and personal characteristics. *BMC Medical Education*, *16* (1), 1-11.

JACOBS, J.C., WILSCHUT, J., VAN DER VLEUTEN, C., SCHEELE, F., CROISET, G., & KUSURKAR, R.A. (2020). An international study on teachers' conceptions of learning and teaching and corresponding teacher profiles. *Medical Teacher*, *42* (9), 1000-1004.

JUDD, C.M., JESSOR, R., & DONOVAN, J.E. (1986). Structural equation models and personality research. *Journal of Personality*, *54* (1), 149-198. doi.org/10.1111/j.1467-6494.1986.tb00392.x

KEMBER, D. (2009). Promoting student-centred forms of learning across an entire university. *Higher Education*, *58* (1), 1-13.

KIM, H.-R., SONG, Y., LINDQUIST, R., & KANG, H.-Y. (2016). Effects of team-based learning on problem-solving, knowledge and clinical performance of Korean nursing students. *Nurse Education Today*, *38*, 115-118.

KLINE, R.B. (2011). *Principles and practice of structural equation modeling*. New York, NY: Guilford.

LEYS, C., KLEIN, O., DOMINICY, Y., & LEY, C. (2018). Detecting multivariate outliers: Use a robust variant of the Mahalanobis distance. *Journal of Experimental Social Psychology*, *74*, 150-156.

LI, Y., WANG, X., ZHU, X., ZHU, Y., & SUN, J. (2019). Effectiveness of problem-based learning on the professional communication competencies of nursing students and nurses: A systematic review. *Nurse Education in Practice*, *37*, 45-55.

MATSUNAGA, M. (2010). How to factor-analyze your data right: Do's, Don'ts, and How-To's. *International Journal of Psychological Research*, *3* (1), 97-110.

McLEAN, M., & GIBBS, T. (2010). Twelve tips to designing and implementing a learner-centred curriculum: Prevention is better than cure. *Medical Teacher*, *32* (3), 225-230.

McLEAN, S.F. (2016). Case-based learning and its application in medical and health-care fields: A review of worldwide literature. *Journal of Medical Education and Curricular Development*, *3*, S20377.

MICHAELSEN, L.K., SWEET, M., & PARMELEE, D.X. (2011). *Team-based learning: Small group learning's next big step*. John Wiley & Sons.

MORIN, A.J.S., MARSH, H.W., & NAGENGAST, B. (2013). Exploratory structural equation modeling. In *Structural equation modeling: A second course*. Information Age Publishing.

MUTHÉN, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika*, *49* (1), 115-132.

NORTON, L., RICHARDSON, T.E., HARTLEY, J., NEWSTEAD, S., & MAYES, J. (2005). Teachers' beliefs and intentions concerning teaching in higher education. *Higher Education*, *50* (4), 537-571.

NUNDY, S., KAKAR, A., & BHUTTA, Z.A. (2022). The why and how of problem-based learning? In *How to Practice Academic Medicine and Publish from Developing Countries?* Springer.

NUNNALLY, J.C. (1978). *Psychometric theory 2nd edition*. New York: McGraw.

PACIFICO, J.L., VAN MOOK, W., DONKERS, J., JACOBS, J.C., VAN DER VLEUTEN, C., & HEENEMAN, S. (2021). Extending the use of the Conceptions of Learning and Teaching (COLT) instrument to the postgraduate setting. *BMC Medical Education*, *21* (1), 1-6.

PAJARES, M.F. (1992). Teachers' beliefs and educational research: Cleaning up a messy construct. *Review of Educational Research*, *62* (3), 307-332.

POSTAREFF, L. (2007). *Teaching in higher education: From content-focused to learning-focused approaches to teaching*.

POSTAREFF, L., LINDBLOM-YLÄNNE, S., & NEVGI, A. (2008). A follow-up study of the effect of pedagogical training on teaching in higher education. *Higher Education*, *56* (1), 29-43.

PRATT, D.D., & COLLINS, J.B. (2001). *Teaching Perspectives Inventory*. Retrieved November, 10, 2012.

ROTTHOFF, T., BAEHRING, T., DAVID, D.M., BARTNICK, C., LINDE, F., WILLERS, R., SCHÄFER, R.D., & SCHERBAUM, W.A. (2011). The value of training in communication skills for continuing medical education. *Patient Education and Counseling*, *84* (2), 170-175.

SADLER, I. (2012). The challenges for new academics in adopting student-centred approaches to teaching. *Studies in Higher Education*, *37* (6), 731-745.

SNELGROVE, H., FAMILIARI, G., GALLO, P., GAUDIO, E., LENZI, A., ZIPARO, V., & FRATI, L. (2009). The challenge of reform: 10 years of curricula change in Italian medical schools. *Medical Teacher*, *31* (12), 1047-1055.

SULAIMAN, N.A., SHAHIMI, S., & ZAKARIA, Z. (2021). Student-centred, lecturer-centred, and hybrid teaching methods: Impacts on accounting students' technical and soft skills. *Asian Journal of Accounting Perspectives*, *14* (2), 67-92.

SURESH, K.P., & CHANDRASHEKARA, S. (2012). Sample size estimation and power analysis for clinical research studies. *Journal of Human Reproductive Sciences*, *5* (1), 7.

TRIGWELL, K., & PROSSER, M. (1996). Congruence between intention and strategy in university science teachers' approaches to teaching. *Higher Education*, *32* (1), 77-87.

TRIGWELL, K., PROSSER, M., & WATERHOUSE, F. (1999). Relations between teachers' approaches to teaching and students' approaches to learning. *Higher Education*, *37* (1), 57-70.

ULFA, Y., IGARASHI, Y., TAKAHATA, K., SHISHIDO, E., & HORIUCHI, S. (2021). A comparison of team-based learning and lecture-based learning on clinical reasoning and classroom engagement: A cluster randomized controlled trial. *BMC Medical Education*, *21* (1), 1-11.

WALD, H.S., McFARLAND, J., & MARKOVINA, I. (2019). Medical humanities in medical education and practice. *Medical Teacher*, *41* (5), 492-496.

XHOMARA, N. (2022). Critical thinking: Student-centred teaching approach and personalised learning, as well as previous education achievements, contribute to critical thinking skills of students. *International Journal of Learning and Change*, *14* (1), 101-120.

ZUMBO, B.D., GADERMANN, A.M., & ZEISSER, C. (2007). Ordinal versions of coefficients alpha and theta for Likert rating scales. *Journal of Modern Applied Statistical Methods*, *6* (1), 4.

Organizational identification: Validation of the Italian scale in the healthcare context

Andrea Caputo[1], Paola Gatti[2], Amelia Manuti[3], Claudio Giovanni Cortese[1]

[1] Department of Psychology, University of Turin, Italy
[2] Department of Psychology, University of Milano-Bicocca, Italy
[3] Department of Education, Psychology, Communication, University of Bari, Italy

andrea.caputo@unito.it

✦ *ABSTRACT*. Questo studio propone la validazione della versione italiana della scala di organizational identification proposta da Mael e Ashforth (1992), adattata al contesto sanitario. L'identificazione organizzativa riguarda la percezione dei lavoratori di sentirsi uniti alla propria azienda e questo sentimento può favorire comportamenti di cittadinanza organizzativa e proteggere dalle intenzioni di turnover, aspetti cruciali soprattutto per le organizzazioni sanitarie. Lo studio ha coinvolto 1505 infermieri del settore pubblico. I risultati confermano la versione italiana come uno strumento valido e affidabile nella valutazione dell'identificazione degli infermieri con la propria azienda sanitaria. L'uso di questo strumento può contribuire al miglioramento degli indici di benessere organizzativo e di retention degli infermieri nelle aziende sanitarie.

✦ *SUMMARY. Organizational identification is related to employees' perception of oneness with their workplace. Being identified with one's organization could promote organizational citizenship behaviors and protect from turnover, very important performance outcomes especially in healthcare organizations. Mael and Ashforth's (1992) scale is one of the most used tools in literature, and this study proposes the validation of the Italian version within the healthcare context. 1505 nurses working in public sector were involved. Confirmatory factor analyses, multigroup and invariance tests, and reliability analyses were performed. Convergent and divergent validity were tested with correlational analyses. Results confirm the Italian version as a valid and reliable tool, facilitating the evaluation of nurses' identification with their healthcare organization. This validation allows enhancing understanding of organizational dynamics within healthcare contexts, ultimately contributing to the development of efficient management strategies and to the improvement of outcomes for both staff and patients.*

INTRODUCTION

Several studies have explored the role played by organizational identification in shaping employees' work engagement, job satisfaction and turnover intentions or organizational citizenship (Karanika-Murray, Duncan, Pontes & Griffiths, 2015; Urbini, Chirumbolo, Caracuzzo & Callea, 2023). Organizational identification arises from the wider construct of social identification (Mael & Ashforth, 1992). Mael and Ashforth (1992) defined organizational identification as "the perception of oneness with or belongingness to an organization, where the individual defines him or herself in terms of the organization(s) in which he or she is a member" (p. 104). Organizational identification is related to social identity theory (Tajfel & Turner, 1985), since individuals tend to classify both themselves and others into various social groups, including the belongingness to specific organizations. According to the process of social identification, individual perceives themselves as a member of a specific group, could it be a football team (e.g., "we" won the match) or an organization. Moreover, organizational identification has been described as a cognitive construct, and as a relational and comparative construct since individuals define themselves in terms of their membership to a specific organization with respect to another.

Mael and Ashforth (1992) distinguished organizational identification from other comparable constructs, such as organizational commitment and professional identification. Firstly, commitment involves an individual's acceptance of the organization's goals and values, willingness to exert effort on its behalf, and desire to maintain membership. Unlike organizational identification, which entails a sense of belonging to a specific organization, commitment does not necessarily include perceiving a collective destiny with that organization. Secondly, professional identification pertains to how individuals perceive themselves embodying the prototypical traits of a certain profession (Mao, Lu, Lin & He, 2021), thus not exclusively tied to a single organization, as the profession could be practiced across various organizations.

As regards antecedents and outcomes, a meta-analytic study (Lee, Park & Koo, 2015) showed that organizational identification is significantly associated with attitudes (such as job involvement and satisfaction) and behaviors (i.e., in-role and extra-role performance). Specifically, it could be considered a predictor for general attitude and behavior. Furthermore, it is related to a sense of pride in being part of a particular organization. Organizational studies defined organizational identification as a mediator between antecedents and outcomes or as an outcome itself (Riketta, 2005).

Identification with healthcare organizations

Healthcare organizations are a peculiar professional context, due to the high demands the workforce have to face, the high rate of turnover and the high quality of care expected (Chen, Yu, Hsu, Lin & Lou, 2013; Katrinli, Atabay, Gunay & Guneri, 2008). Specifically, nurses represent the segment of the workforce with a steadily increasing turnover rates (Hayes et al., 2012). In Italy, between 2010 and 2019, there has been a rising detrimental turnover trend, with nurses not being adequately replaced (FNOPI, 2022; https://www.fnopi.it/wp-content/uploads/2022/10/AGENAS-personale_ssn_2022.pdf). Among the constructs that can protect nurses from exhaustion, organizational identification plays an important role in increasing nurses' performance and good contextual resources, since it is related to employees' perception of oneness with their organization (Katrinli et al., 2008). Studies within the healthcare context highlight that organizational identification could be affected by the quality of the relationship between nurses and their nurse leaders, and protect from turnover intentions (Katrinli et al., 2008). Thus, revisiting psychometrical properties of Organizational Identification Scale to highlight its inherent value is important, also to improve research within this specific target population.

Quantitative studies have employed a range of scales to evaluate organizational identification, yet the most prevalent in organizational research is the scale developed by Mael and Ashforth (1992), as highlighted by prior meta-analytic research (Riketta, 2005). Despite this, scale validation studies are surprisingly scarce, and the few that do exist tend to focus on generic organizations rather than being tailored to specific contexts. Indeed, in Italy, the psychometric properties of Mael and Ashforth's scale were tested on a sample of employees from micro and little organizations (Manuti & Bosco, 2012). This gap in the literature suggests a need for more nuanced validation efforts that consider the unique characteristics and dynamics of different organizational environments.

The present study

The present study aims to validate the Italian version of Mael and Ashforth's *Organizational Identification Scale* within a sample of nurses working in public sector hospitals. Thus, the process of developing the Italian version of the original *Organizational Identification Scale* was realized starting from the original items in Mael and Ashforth's scale (1992) to include in the wording of each item the specific mention of the healthcare organization in which the participant works. We then tested measurement invariance and performed a multigroup analysis. Additionally, we explored the convergent and divergent validity of the instrument to ensure its robustness and applicability in this specific context.

METHOD

Participants and procedure

Italian nurses and nurse leaders employed in 4 major hospitals belonging to the same healthcare organization in northwestern Italy participated in this study by completing paper and pencil questionnaires. This research came from a broader project titled "Feeling like a leader", which aimed to explore the dynamics of leadership relationships between nurse leaders and the nurses in their respective working groups. To safeguard participants' privacy, alphanumeric codes were generated to match nurse leaders with their respective follower groups, while ensuring confidentiality. Participants were briefed on the process through invitation letters and accompanying information sheets. Data collection started after approval from the Director of the Directorate of Health Professions and the nurse leaders of the targeted organization, as well as clearance from the Bio-Ethics Committee of the University of Turin (Approval letter, Prot. No. 55631, dated 01.02.2019). Nurse leaders received email invitations along with detailed research information. Upon their agreement to participate, paper questionnaires were personally delivered and collected by administrators. The study included the entire population of nurses from the targeted organizations, totaling 2664 individuals. A criterion for inclusion was the completion of at least 61% of all questionnaire items. Setting a threshold for the minimum percentage of completed items necessary for a respondent's data to be included in the analysis is a common practice. This

threshold can vary, but commonly used benchmarks range from 60% to 80% (e.g., Hox & De Leeuw, 1994). We selected a threshold slightly above the minimum to avoid employing an overly rigid and restrictive criterion. Following data cleaning, the final sample comprised 1550 nurses, representing 58.2% of the total population. The sample consists of 82.6% of women and 17.4% of men, with an average age of 43.4 years old ($SD = 9.2$). Regarding departmental distribution, 35.3% of nurses worked in medicine wards, 29.6% in surgery, 15% in emergency, and 20.1% in pediatrics. On average, nurses had been employed in their current hospital, i.e., tenure in the hospital, for 17.5 years ($SD = 9.9$), with a tenure within their specific ward averaging 11.5 years ($SD = 8.3$).

Measures

– *Organizational identification* was measured with the Italian version of Mael and Ashforth (1992) scale. The development of the Italian version followed a back-translation procedure (Brislin, 1970), starting from the original items in Mael and Ashforth's scale (1992). Instead of referring to a specific school (as the original authors did) or a specific firm labeled as "the organisation I work for" (as the Italian authors did in their version for generic organizations; Manuti & Bosco, 2012, p. 897), we tailored each item to specifically mention the name of the healthcare organization where the nurses work (see Appendix). This approach references the larger healthcare organization, which may encompass multiple hospitals, while still referring to the same overall entity, as healthcare organizations typically consist of various specialized hospitals. Initially, the original 6 items from Mael and Ashforth's (1992) scale were translated into Italian by the authors and then blindly translated back into English by a native speaker. Any minor discrepancies from the original wording were addressed to ensure the items were easily comprehensible for participants, who were asked to answer using a 6-point Likert scale (from 1 = Strongly disagree to 6 = Strongly agree), in order to avoid people choosing the central point, forcing them to take sides (Preston & Colman, 2000). Appendix shows items in both English and Italian versions.
– *Organizational tenure* was assessed with a single item asking participants "How many years have you been working at this hospital?".

- *Professional identification* was assessed with the adapted version of the 4-item scale by Ostermeier (2018), with a 5-point Likert scale (from 1 = Strongly disagree to 5 = Strongly agree). An example item is "My profession has a clear and unique vision". McDonald's ω = .82.
- *Job satisfaction* was assessed with a 5-item scale from the *Copenhagen Psychosocial Questionnaire* (*COPSOQ*; Pejtersen et al., 2010), with a 5-point Likert scale (from 1 = Strongly disagree to 7 = Strongly agree). An example item is "How satisfied are you with your work as a whole, taking into consideration each element?". McDonald's ω = .87.
- *Work engagement* was assessed with a 9-item scale, i.e. the short version of the *Utrecht Work Engagement Scale* (*UWES*; Schaufeli et al., 2006) in its Italian validated version (Balducci. Fraccaroli & Schaufeli, 2010), with a 7-point Likert scale (from 0 = Never to 6 = Always). An example item is "At my work, I am bursting with energy". McDonald's ω = .91.
- *Emotional exhaustion* was assessed with an 8-item scale from Oldenburg Burnout Inventory (*OLBI*; Demerouti, Mostert & Bakker, 2010), with a 4-point Likert scale (from 1 = Strongly disagree to 4 = Strongly agree). An example item is "During my work, I often feel emotionally drained". McDonald's ω = .77.

Data analyses

Analyses were conducted using SPSS for descriptive statistics, reliability, and correlational analyses, and Mplus for confirmatory factor, multigroup, and invariance analyses. Univariate and multivariate skewness and kurtosis were examined to assess the distributional characteristics. For multivariate normality assessment, Mardia's coefficients were computed using a web tool available at the following link: https://webpower.psychstat.org/models/kurtosis.

Confirmatory factor analyses (CFAs) employing the maximum likelihood (ML) estimator were conducted to assess the model fit of the *Organizational Identification Scale*. The adequacy of model fit was evaluated using established thresholds for favorable model fit: Comparative Fit Index (CFI) and Tucker-Lewis Index (TLI) values >.90/.95, Root Mean Square Error of Approximation (RMSEA) values <.05/.08, and Standardized Root Mean Square Residual (SRMR) values <.08 (Little, 2013).

RESULTS

Descriptive statistics and reliability

Regarding univariate skewness, as detailed in Table 1, the six items measuring organizational identification demonstrated a normal distribution, with skewness values falling below ±2 and kurtosis values below ±7.

The results indicated a multivariate skewness coefficient of .27 (*p*<.001) and a multivariate kurtosis coefficient of 2.72 (*p*<.001). Given that the items related to organizational identification exhibited Mardia's coefficients below ±3 (Bandalos & Finney, 2010), they can be considered to conform to a normally distributed data pattern.

Reliability analyses evaluate the internal consistency of the scale and explore to what degree the scores are free from random measurement error. Cronbach's α, McDonald's ω, the average variance extracted (AVE) and the composite reliability (CR) were assessed (Fornell & Larcker, 1981). Acceptable values are above .70 for both Cronbach's α, McDonald's ω and CR, and above .50 for AVE; moreover, AVE should be smaller than CR (Hair, Black, Babin & Anderson, 2010). Table 2 shows reliability indices of the *Organizational Identification Scale*. All scores above the thresholds indicates a good reliability of the scale.

Confirmatory factor analysis, multigroup, and invariance tests

The CFA on the Italian translation of the *Organizational Commitment Scale* showed the following model fit: $\chi^2_{(9)}$= 235.501 (*p*<.001); CFI = .95; TLI = .91; RMSEA = .13 [.116; .145]; SRMR = .04. Table 3 shows standardized factor loadings.

Subsequently, we examined measurement invariance by comparing two subsamples based on organizational tenure. To accomplish this, the organizational tenure variable was dichotomized using the median, resulting in two groups: nurses with tenures up to 17 years and those with tenures ranging from 18 to 43 years. First, we conducted a CFA for each of the two groups separately. Subsequently, we explored measurement invariance according to the four levels delineated in the literature (Meredith, 1993). The first level is configural invariance, which assessed a model with no invariance constraints, serving as a baseline comparison.

Table 1 – Descriptive statistics of Organizational Identification Scale, Italian version

Item	Mean	*SD*	Skewness		Kurtosis	
			Stats	*SE*	Stats	*SE*
Orgid_1	3.46	1.550	−.113	.064	−1.063	.127
Orgid_2	3.64	1.445	−.264	.063	−.824	.127
Orgid_3	4.43	1.433	−.849	.063	−.086	.127
Orgid_4	3.92	1.509	−.466	.064	−.753	.127
Orgid_5	4.07	1.493	−.551	.064	−.665	.127
Orgid_6	4.10	1.514	−.525	.064	−.685	.127

Legenda. SD = Standard deviation; *SE* = Standard error.

Table 2 – Reliability indices of Organizational Identification Scale, Italian version

	Cronbach's α	**McDonald's ω**	**AVE**	**CR**
Organizational identification	.88	.88	.56	.88

Legenda. AVE = Average variance extracted; CR = Composite reliability.

Table 3 – Factor loadings from CFA (ML estimator)

Items	Standardized estimates	*t*-value	*p*
Orgid_1	.70	46.92	<.001
Orgid_2	.64	36.91	<.001
Orgid_3	.73	51.69	<.001
Orgid_4	.84	84.92	<.001
Orgid_5	.86	94.74	<.001
Orgid_6	.68	42.78	<.001

The second level, weak invariance, tested the equivalence of factor loadings across groups. Achieving weak invariance suggests that factor loadings of items are consistent across groups. Moving on, the third level, strong invariance, entailed constraining item intercepts. If attained, this level permits comparisons of means across groups. Finally, the fourth level, strict invariance, was examined for invariance in error variances. Table 4 shows the comparisons of CFAs for the two separate groups and of the models for measurement invariance.

Results of the CFA performed in both groups divided by tenure (i.e., group 1 = working up to 17 years in the organization, and group 2 = working between 18 and 43 years in the organization) indicated that the models exhibited acceptable fit to the data in both samples.

Regarding the invariance test, the configural model demonstrated a good fit to the data, implying that the model adequately captured the data from both samples without imposing additional invariance constraints. Then, metrical (weak) model invariance was also supported, since the difference in fit between the weak and configural models was not statistically significant. Also, the scalar (strong) model displayed a good fit compared to the weak model. Finally, the strict model exhibited a slightly worse fit compared to the scalar (strong) model based on the chi-squared difference. However, the changes in RMSEA and CFI were below the recommended thresholds of .015 and .01, thereby supporting strong measurement invariance and enabling comparisons of means between the two samples.

Convergent and divergent validity

Convergent validity assesses the degree to which a measure correlates positively with other measures that it theoretically should correlate with, thereby confirming expected relationships between related constructs. Conversely, divergent validity ensures that the measure is distinct from unrelated constructs and accurately captures the intended construct without measuring unintended aspects. Convergent and divergent validity were assessed correlating organizational identification with the organizational tenure in hospital, constructs related to other kind of identification (i.e. professional identification), and some classical organizational outcomes (i.e. job satisfaction,

work engagement, emotional exhaustion). Table 5 shows results of the correlation analysis.

Results of correlations table confirms the relatedness of organizational identification with organizational tenure, in line with literature (Chen et al., 2013), it is also positively related to constructs of identification and group cohesiveness, furthermore it shows convergent validity with some organizational wellbeing outcomes, while showing divergent validity with emotional exhaustion.

DISCUSSION

This study highlights the effectiveness of the Italian version of the *Organizational Identification Scale* in measuring how nurses identify with their employing healthcare organization. This tool holds significant potential for advancing research within the healthcare sector. By assessing organizational identification, which is often investigated as a moderator or mediator in research frameworks, it provides deeper insights into organizational dynamics that influence both performance and wellbeing (Lee et al., 2015).

Limitation and future studies

One initial limitation concerns the cross-sectional design of the study and reliance on self-reported data. Moreover, this study is focused only on nurses. Additionally, nurses who participated in the study were employed in hospitals located within the same city and only in a public organization. Therefore, achieving a robust definition of the constructs will necessitate integration with additional investigations conducted nationwide and involving nurses working in the private sector. Future studies could ensure the reliability of the instrument with multigroup analyses referring to a sample of nurses working in private and public organizations; also, future research could use this tool to evaluate the extent to which nurses would identify with their own healthcare organization (whether public or private) from a longitudinal perspective, to assess changes of identification over time, and connecting this evidence to contextual events, also in relation to significant outcomes for nurses, such as job satisfaction (Gatti et al., 2020). Finally, future validation efforts could involve a broader range of healthcare staff.

Table 4 – Results of CFA and multigroup invariance tests (ML estimator)

Model	χ^2	*df*	p	$\Delta\chi^2$	Δdf	*p*	CFI	TLI	RMSEA (90% CI)	SRMR	ΔCFI
					Single groups models						
Tenure 1*	104.169	9	<.001				.961	.935	.122 [.101; .143]	.033	
Tenure 2**	137.976	9	<.001				.925	.876	.144 [.123; .165]	.043	
					Multiple groups invariance						
Configural	242.146	18	<.001				.946	.910	.133 [.118; .148]	.038	
Metric (weak)	250.200	23	<.001	8.05	5	.153	.945	.929	.118 [.105; .132]	.045	.001
Scalar (strong)	286.111	29	<.001	35.91	6	<.001	.938	.936	.112 [.101; .124]	.066	.007
Strict	323.986	35	<.001	323.99	12	<.001	.931	.940	.108 [.098; .119]	.059	.007

Legenda. df = degree of freedom; CFI = Comparative Fit Index; TLI = Tucker-Lewis Index; RMSEA = Root Mean Square Error of Approximation; SRMR = Standardized Root Mean Square Residual.

Note. * Organizational tenure up to 17 years, *N* = 724; ** Organizational tenure between 18 and 43 years, *N* = 699.

Table 5 – Correlation table

	1.	2.	3.	4.	5.
1. Organizational identification	–				
2. Organizational tenure	.136***	–			
3. Professional identification	.177***	.012	–		
4. Job Satisfaction	.312***	.097***	.136***	–	
5. Work engagement	.372***	.035	.193***	.581***	–
6. Emotional exhaustion	–.143***	–.003	–.140***	–.460***	–.474***

*** *p*<.001

CONCLUSION

This study contributes to the existing literature by validating the efficacy of the Italian version of the *Organizational Identification Scale* among nurses, a workforce that interacts closely with patients and contends with numerous daily demands. Given the significance of organizational identification, human resource management within healthcare institutions, including managers such as nurse managers or head physicians, should prioritize efforts to enhance it. Therefore, organizational wellbeing and citizenship behaviors among nurses could be promoted, thereby improving both the employer brand to strengthen employee retention (Caputo, Molino, Cerato & Cortese, 2023) and the quality of patient care, mitigating challenges associated with workforce turnover (Lee et al., 2015).

References

BALDUCCI, C., FRACCAROLI, F. & SCHAUFELI, W. B. (2010). Psychometric properties of the Italian version of the Utrecht Work Engagement Scale (UWES-9). *European Journal of Psychological Assessment, 26* (2), 143-149. doi.org/10.1027/1015-5759/a000020

BANDALOS, D.L. & FINNEY, S.J. (2010). Factor analysis: Exploratory and confirmatory. In *The reviewer's guide to quantitative methods in the social sciences*. New York, NY: Routledge.

BRISLIN, R.W. (1970). Back-translation for cross-cultural research. *Journal of Cross-Cultural Psychology*, *1* (3), 185-216. doi.org/10.1177/135910457000100301

CAPUTO, A., MOLINO, M., CERATO, B. & CORTESE, C.G. (2023). Employer attractiveness: Two instruments to measure employer branding and reputation. *SAGE Open*, *13* (3), 21582440231192188. doi.org/10.1177/21582440231192188

CHEN, S.H., YU, H.Y., HSU, H.Y., LIN, F.C. & LOU, J.H. (2013). Organisational support, organisational identification and organisational citizenship behaviour among male nurses. *Journal of Nursing Management*, *21* (8), 1072-1082. doi.org/10.1111/j.1365-2834.2012.01449.x

DEMEROUTI, E., MOSTERT, K. & BAKKER, A.B. (2010). Burnout and work engagement: A thorough investigation of the independency of both constructs. *Journal of Occupational Health Psychology*, *15* (3), 209-222. doi.org/10.1037/a0019408

FORNELL, C. & LARCKER, D.F. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of Marketing Research*, *18* (1), 39-50. doi.org/10.1177/002224378101800104

GATTI, P., PEPE, E., CAPUTO, A., CLARI, M., DIMONTE, V., HALL, R. & CORTESE, C. (2020). Job satisfaction in a sample of nurses: A multilevel focus on work team variability about the head nurse's transformational leadership. *Electronic Journal of Applied Statistical Analysis*, *13* (3), 713-738. doi.org/10.1285/i20705948v13n3p713

HAIR, J.F., BLACK, W.C., BABIN, B.J. & ANDERSON, R.E. (2010). *Multivariate Data Analysis (7th ed.)*. Englewood Cliffs, NJ, USA: Prentice Hall.

HAYES, L.J., O'BRIEN-PALLAS, L., DUFFIELD, C., SHAMIAN, J., BUCHAN, J., HUGHES, F., LASCHINGER H.K.S., NORTH, N. & STONE, P.W. (2012). Nurse turnover: A literature review – an update. *International Journal of Nursing Studies*, *49* (7), 887-905. doi.org/10.1016/j.ijnurstu.2011.10.001

HOX, J.J. & DE LEEUW, E.D. (1994). A comparison of nonresponse in mail, telephone, and face-to-face surveys: Applying multilevel modeling to meta-analysis. *Quality and Quantity*, *28* (4), 329-344.

IBM CORP. (Released 2023). *IBM SPSS Statistics for Windows*, *Version 29.0.2.0*. Armonk, NY: IBM Corp.

KARANIKA-MURRAY, M., DUNCAN, N., PONTES, H.M. & GRIFFITHS, M.D. (2015). Organizational identification, work engagement, and job satisfaction. *Journal of Managerial Psychology*, *30* (8), 1019-1033. doi.org/10.1108/JMP-11-2013-0359

KATRINLI, A., ATABAY, G., GUNAY, G. & GUNERI, B. (2008). Leader-member exchange, organizational identification and the mediating role of job involvement for nurses. *Journal of*

Advanced Nursing, *64* (4), 354-362. doi.org/10.1111/j.1365-2648.2008.04809.x

LEE, E.-S., PARK, T.-Y. & KOO, B. (2015). Identifying organizational identification as a basis for attitudes and behaviors: A meta-analytic review. *Psychological Bulletin*, *141* (5), 1049-1080. doi.org/10.1037/bul0000012

LITTLE, T.D. (2013). *Longitudinal structural equation modeling*. New York, NY: Guilford.

MAEL, F. & ASHFORTH, B.E. (1992). Alumni and their alma mater: A partial test of the reformulated model of organizational identification. *Journal of Organizational Behavior*, *13* (2), 103-123. doi.org/10.1002/job.4030130202

MANUTI, A. & BOSCO, A. (2012). Identificazione organizzativa: Un contributo alla verifica delle proprietà psicometriche di due strumenti di misura. *Giornale Italiano di Psicologia*, *39* (4), 881-902. doi: 10.1421/73147

MAO, A., LU, S.E., LIN, Y. & HE, M. (2021). A scoping review on the influencing factors and development process of professional identity among nursing students and nurses. *Journal of Professional Nursing*, *37* (2), 391-398. doi.org/10.1016/j.profnurs.2020.04.018

MEREDITH, W. (1993). Measurement invariance, factor analysis, and factorial invariance. *Psychometrika*, *58*, 525-543. doi.org/10.1007/BF02294825

MUTHÉN, L.K. & MUTHÉN, B.O. (1998-2011). *Mplus User's Guide. Sixth Edition*. Los Angeles, CA: Muthén & Muthén.

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

TAJFEL, H. & TURNER, J.C. (1985). The social identity theory of intergroup behavior. In S. Worchel & W.G. Austin (Eds.), *Psychology of Intergroup Relations*. Nelson-Hall, Chicago.

URBINI, F., CHIRUMBOLO, A., CARACUZZO, E. & CALLEA, A. (2023). Job satisfaction and OCBs: What's new? The mediating role of Organizational Identification. *International Journal of Organizational Analysis*, *31* (5), 1748-1762. doi.org/10.1108/IJOA-07-2021-2865

ZHANG, Z. & YUAN, K.-H. (Eds.) (2018). *Practical statistical power analysis using Webpower and R*. Granger, IN: ISDSA Press.

APPENDIX

English and Italian version of Organizational Identification Scale

English version (Mael & Ashforth, 1992)	Italian version
When someone criticizes (name of school), it feels like a personal insult	Quando qualcuno critica l'azienda ospedaliera in cui lavoro, mi sento insultato personalmente
I am very interested in what others think about (name of school)	Sono molto interessato a sapere quello che gli altri pensano dell'azienda ospedaliera in cui lavoro
When I talk about this school, I usually say 'we' rather than 'they'	Quando parlo dell'azienda ospedaliera in cui lavoro di solito dico "noi" piuttosto che "loro"
This school's successes are my successes	I successi dell'azienda ospedaliera in cui lavoro sono i miei successi
When someone praises this school, it feels like a personal compliment	Quando qualcuno elogia l'azienda ospedaliera in cui lavoro, è come se mi facesse un complimento personale
If a story in the media criticized the school, I would feel embarrassed	Se una notizia nei mass-media criticasse l'azienda ospedaliera in cui lavoro, mi sentirei a disagio