Validation of the German group development (GD) questionnaire

Jan-Paul Leuteritz [1] · Jose Navarro [2] · Jan Philipp Czakert [2] · Rita Berger [2] 🔵

Abstract

Group Development (GD) is an important variable when researching and evaluating what makes teams successful. We analyzed the psychometric properties of the originally Spanish GD questionnaire with German participants. 501 team members and 104 team leaders, 18 to 65 years old, from a German research organization answered an online survey composed of the GD questionnaire and items related to other group processes of democracy, mutual trust, team spirit, and interest in the team's tasks. Results confirmed the unidimensional factor structure of the translated Spanish version for the German GD construct. Internal consistency, convergent and discriminant validity were good. The German GD correlated as expected to other constructs, and it showed concurrent validity with respect to the *team members' motivation and interest in team tasks* ($r = .79$, $p < .01$). We recommend using the GD in German samples to measure team processes that are highly relevant for team effectiveness.

Keywords Validation · Questionnaire · Team · Workgroup · Group development · Confirmatory factor analysis

Organizations rely increasingly on teams. Thus, assessing group development is a relevant aspect of human resource development. Group development models (Tuckman and Jensen 1977; Wheelan and Hochberger 1996) concentrate on developmental stages or a group's life cycle. Since they suppose linear development (Miller 2003), they fail to explain what occurs after a team has reached the final stage. We expect this to cause ceiling effects or eliminate variance in teams that have not been set up recently. An alternative would be measuring a continuous variable; researchers have often chosen group cohesion to assess a group's capacity of creating emergence or working as a team. However, this approach has been criticized (Hogg 1993), mostly due to the different operationalizations of cohesion.

As valid alternative instruments for the continuous measurement of a group's development were missing, Meneses et al. (2008) created the group development (GD) questionnaire. The GD instrument consists of 8 items that form a single factor. It assesses group development as the degree to which a group shows the following characteristics: 1) interpersonal relationships among its members; 2) identification of the members with their group; 3) coordination of behaviors, resources and technologies; and 4) members being dedicated to achieving group goals (Navarro et al. 2015). The final GD questionnaire resulted from validation studies in Spain, Brazil, and Venezuela (Navarro et al. 2015).

To create the GD, Meneses et al. (2008) analyzed the literature and identified two main approaches: on the one hand, the stage approach of a developing group (Wheelan and Hochberger 1996) and, on the other hand, three families of continuous group measurements: groupness (Arrow et al. 2000), entitativity (Hamilton and Sherman 1996), and groupality (Roca Cortés 2001).

Among the latter three construct families, they found similarities in 8 dimensions (interrelationship, shared goals, identification with the group, group coordination, shared results, task interdependence, social value of the task, and orientation to group goals). Based on these dimensions, the GD was developed, eventually resulting in a measurement representing the aforementioned 4 characteristics of well-developed groups.

To date, most research on groups has been based on IPO (Input-Process-Output) models (Marks et al. 2001). Despite the criticism regarding IPO frameworks (Ilgen et al. 2005), it may be relevant for future users of the GD to know whether it represents a group process or rather an emergent state. Based on the instrument itself, as well as on the theoretical ingredients used to create the GD, we argue that it measures a group process.

✉ Rita Berger
  ritaberger@ub.edu

1 Fraunhofer-Institute for Industrial Engineering (IAO), Berlin, Germany

2 Departamento de Psicología Social y Psicología Cuantitativa, Universitat de Barcelona, Passeig de la Vall d'Hebron, 171, E-08035 Barcelona, Spain

The revised concept of groupness, as presented by McGrath (1984) was originally based on group fuzziness, thus relating to the work group's characteristics (e.g., number of members) and relationship patterns. Groupness is a "fundamental process for the existence of a group and one that explains the extent to which a set of people can be characterized by specific variables that enable it to be perceived as a group or an aggregate" (Meneses et al. 2008, p. 495). The criteria proposed by Arrow et al. (2000) also resemble processes or activities (e.g., coordination of behaviors) rather than cognitive or emotional states.

Entitativity refers "specifically to the degree to which a group really exists, that is, the extent to which it exists as an entity" (Meneses et al. 2008, p. 498). Entitativity mainly represents perceived unity, from the perspective of insiders or outsiders of the respective group. Hamilton et al. (2013) operationalized entitativity through organization and structure among the members, which resembles properties (i.e., emergent states) of the group rather than processes; nonetheless, processes are directly linked to these properties (leadership, performance). Lickel et al. (2000) proposed 5 variables 2000): (a) interaction, (b) common goals, (c) common results, (d) similarity among members of the group, and (e) importance of the group for its members. These variables represent structure, process, and results, which seems to intermingle states and processes.

The concept of groupality originates from Soviet psychology and was adapted by Roca Cortés (2001) in the form of "characteristic attributes of a group that are present in personal interaction and that change both in quality and intensity as a function of time and group activity" (Meneses et al. 2008, p. 503). Roca Cortés (2001) proposed the following 5 dimensions: (a) social value of activity content, (b) communication and interpersonal relationships, (c) group goals, (d) leadership and management, and (e) group organization and group influence over its members. This concept is focused on the interaction between members and a measurement of processes rather than emergent states.

With the 4 characteristics mentioned above, the GD has the greatest overlap with groupality, and thus should be considered a team process. Each of these characteristics resembles a group process: interpersonal relationships among members and coordination of behaviors, resources, and technologies clearly relate to processes between members. Identification of the members with the group, and the members' orientation toward achieving group goals, represent processes at the individual level.

Navarro et al. (2015) found the GD to correlate to measures of entitativity (Carpenter and Radhakrishnan 2002) at $r = .77$ ($p < .01$) and with group identification (Hogg et al. 1990) at $r = .75$ ($p < .01$). Navarro et al. (2016) found that GD scores were in line with Wheelan's GDQ phases (Wheelan and Hochberger 1996): the correlation between the global GDQ score and the GD was $r = .74$ ($p < .01$), and the correlation between GD score and the GDQ Phase IV score was $r = .79$ ($p < .01$). At GDQ phase IV, the most progressive phase that still comes with a measurement, the group "gets, gives, and uses feedback about its effectiveness and productivity"; it "acts on its decisions"; and it "encourages high performance and quality work" (Wheelan and Hochberger 1996, p. 157). Furthermore, Navarro et al. (2016) reported that the GD correlates with measures of the satisfaction of needs among the group members, $r = .67$ ($p < .01$).

The GD predicts the performance of a group, as measured by a questionnaire based on the criteria by Hackman (1987), and by the indicators *absenteeism* and *order and hygiene* (Navarro et al. 2015). Navarro et al. (2016) evaluated the incremental validity of the GD instrument with respect to the traditional stage-based model of group development as considered by Wheelan and Hochberger's (1996) questionnaire. In comparison with the GDQ, the GD explained, based on Hackman's (1987) criteria (Navarro et al. 2016), additional variance of a group's self-rating of effectiveness. Leuteritz et al. (2017) found that the GD mediated the relationship between transformational leadership and team effectiveness.

To make the instrument available for use in a German-speaking context, and to assess the construct's intercultural relevance, we validated the GD in a German sample. We deemed Germany a good choice because it is Europe's biggest economy.

As a criterion of concurrent validity, we chose the variable *motivation and interest* from the *Team Climate for Learning* (TCl) instrument by Brodbeck et al. (2010). Construct and items share a low level of resemblance with the GD questionnaire. Nevertheless, we assumed that members of a well-developed group would show high engagement in the team's tasks. Regarding convergent validity, we expected a moderate relationship between the TCL dimension *mutual trust* and the GD. We supposed mutual trust to be linked to the GD characteristic of good interpersonal relationships between team members. The TCL dimension *democracy* refers to the absence of dominance by one particular member or leader. Democracy does not have any conceptual overlap with the aspects measured by the GD. Nevertheless, dominant leadership may hamper cooperation (Brodbeck et al. 2010). Consequently, we included democracy as a criterion of discriminant validity.

We also expected the dimension *solidarity* (German: *Zusammenhalt*) in the *Questionnaire on Working in Teams* (German: *Fragebogen zur Arbeit im Team*; F-A-T, Kauffeld and Frieling 2001) to moderately correlate to the GD, since it reflects good personal interrelationships within the team. The F-A-T operationalizes *social reflexivity*, which corresponds to the salience of the group (Kauffeld and Frieling 2001); the GD's precursor constructs of groupness and entitativity also refer to the group's salience. Based on Navarro et al. (2015),

Table 1 Sample

	Team members (N = 501)		Team Leaders (N = 104)	
Age in years: M (SD)	34.3 (11.8)		41.4 (9.5)	
	N	per cent	N	per cent
By gender:				
Male participants	343	68.5%	87	83.7%
Female participants	158	31.5%	17	16.3%
By occupation:				
Researcher	423	84.4%	88	82.2%
Other jobs	78	15.6%	19	17,8

Note. N = Number of individuals

we furthermore expected a German measure of group potency (Moser et al. 2005) to correlate moderately to the GD scores.

Methods

Participants

Team members and their leaders from a German research organization answered an online survey (Table 1). Members and leaders received different questionnaires. Including the leader, each team had at least 4 members (mean team size was 7.9 members). Leaders completed only the GD questionnaire, whereas members also answered items representing group potency, team spirit, democracy, mutual trust, as well as motivation and interest. The category *other jobs* included: administration, IT services, public relations, mechanical work, and facility services.

Measures

Group Development

The GD questionnaire consists of 8 items (Table 2) that come with a 5-point Likert-scale and compose a single factor. Cronbach's α ranged between .70 and .85 in previous validation studies (Navarro et al. 2015). An example item is "We share tools, resources and information." To assess convergent validity, we examined the GD's correlations to similar constructs, using the following validated instruments. Table 3 and Table 4 show the internal consistency coefficients and correlation coefficients from our sample.

Solidarity

We administered all 3 items measuring the factor *solidarity* (German: *Zusammenhalt*) in the F-A-T (Kauffeld and Frieling 2001).

Group Potency

We selected 3 items from the one-dimensional *Scale on work-related collective expectations of efficacy* (Deutsch: *Skala zur arbeitsbezogenen kollektiven Wirksamkeitserwartung*; SABKWSE, Moser et al. 2005).

Democracy, Mutual Trust, and Interest

We used 3 subscales from the *Team Climate for Learning* questionnaire (TCL) by Brodbeck et al. (2010): *democracy* (2 items) for discriminant validity, *mutual trust* (4 items) for

Table 2 GD items

No.	English	German	Spanish
		In meiner Arbeitsgruppe…	
1	We have a usual way of functioning as a group.	haben wir eine Art gefunden wie wir gewöhnlich als Gruppe funktionieren.	Tenemos una forma habitual de funcionar como grupo.
2	We feel we are an important part of this group.	fühlt sich jeder als ein wichtiger Teil dieser Gruppe.	Nos sentimos parte importante de este grupo.
3	All members are consistently relating to one another.	sind wir als Gruppenmitglieder alle ständig miteinander verbunden.	Todos los miembros estamos relacionados constantemente.
4	Members feel committed to the achievement of the group objectives.	fühlen sich die Mitglieder verpflichtet die gesteckten Gruppenziele zu erreichen.	Los miembros se sienten comprometidos en la consecución de las metas del grupo.
5	There is a low interrelation among all members (inverse).	besteht nur wenig Verbundenheit unter den Gruppenmitgliedern.	Hay una baja interrelación entre todos los miembros.
6	We share the same work values.	teilen wir in Bezug auf die Arbeit dieselben Werte.	Compartimos los mismos valores de trabajo.
7	We share tools, resources and information.	teilen wir untereinander Instrumente, Ressourcen und Informationen.	Compartimos herramientas, recursos e información.
8	An essential task is to take care of our own development as a group.	ist es eine grundlegende Aufgabe, uns um unsere eigene Entwicklung als Gruppe zu kümmern.	Una tarea fundamental es cuidar de nuestro propio desarrollo como grupo.

Table 3 Pearson correlations between GD and selected criteria in the members' sample (individual level)

		1	2	3	4	5	6
1	GD	(.87)					
2	TCL Mutual trust	.76**	(.89)				
3	TCL Democracy	.23**	.31**	(.66)			
4	F-A-T Solidarity	−.64**	−.75**	−.30**	(.89)		
5	SABKWSE	.64**	.63**	.24**	−.61**	(.87)	
6	TCL Motivation	.79**	.77**	.23**	−.67**	.71**	(.86)

Note. ** significance at $p < .01$. The principal diagonal (in parentheses) indicates Cronbach's α (members' sample, at individual level)

convergent validity, and the subscale *motivation and interest* (4 items) as external criterion of concurrent validity.

We maintained the original answer formats in all questionnaires named above.

Procedure

Following the guidelines of the International Test Commission (ITC 2017), we applied a back-translation method, to take care of possible cultural or linguistic differences. Tyupa (2011) proposed a framework for back-translation, which we followed as far as applicable, resulting in these steps:

1. Forward translation of the GD from Spanish to German by translator A (female organizational psychologist; native German speaker, fluent in Spanish [C1]) and translator B (male general psychologist; native Spanish speaker, fluent in German [C1]).
2. Revision of the German versions, to identify the items easiest to understand, by 3 employees of German companies.
3. Harmonization of the translated items and the suggestions received from the employees, by translator C (male organizational psychologist; native German speaker, fluent in Spanish [B2]).

Table 4 Pearson correlations among GD and selected criteria in the members' sample (team level)

		1	2	3	4	5	6
1	GD	(.87)					
2	TCL Mutual trust	.79**	(.89)				
3	TCL Democracy	.24**	.30**	(.66)			
4	F-A-T Solidarity	−.69**	−.79**	−.38**	(.89)		
5	SABKWSE	.66**	.60**	.29**	−.59**	(.87)	
6	TCL Motivation	.86**	.82**	.28**	−.67**	.67**	(.86)

Note. ** significance at $p < .01$. The principal diagonal (in parentheses) indicates Cronbach's α (members' sample, at an individual level)

4. Back-translation of the harmonized translation to Spanish by translator D (female Spanish language teacher; German and Spanish bilingually raised native speaker, no access to source text).
5. Review of the back-translation by translator C, in cooperation with B.
6. Adaptation of the German translation, as the back-translation identified discrepancies and possible shifts in meaning or context by Translators A and C.
7. Again back-translation of the new German items by translator D.
8. Review of the second back-translation by translators A and C, who now agreed that sufficient conversion had been achieved.

Since the English version has not been validated yet, it was not used. As in the original, a 5-point Likert scale is used.

We obtained permission for collecting the data from the organization's Human Resource responsible and from the directors of the involved departments. We selected the participants from the organization charts of the involved departments. Each participant received an individual access code through encrypted e-mail. As an incentive, we provided a lottery. The survey was hosted on a European Server. Only one researcher (JPL) had access to the tables that connected the participants' codes to real names; the data files themselves were anonymized. This procedure was applied to both samples: members and leaders.

We used the data provided by the team members for all analyses and performed complementary factor analysis and reliability assessment in the leader data set to broaden the evidence. We ran the Confirmatory Factor Analysis (CFA) in AMOS version 22. To evaluate the Structural Equation Models (SEM), we chose χ^2, χ^2/df ratio, Root Mean Square Error of Approximation (RMSEA) and Tucker-Lewis-Index (TLI) as reference statistics. Following Kenny (2016) we preferred the more conservative TLI to CFI (Comparative Fit Index). We refrained from using Normed Fit Index (NFI), as it does not penalize model complexity (Kenny 2016), and from using Goodness-of-Fit Index (GFI) and Adjusted Goodness-of-Fit Index (AGFI), following Sharma et al. (2005). Nevertheless, to allow for a comparison with the results from the Brazilian sample (Navarro et al. 2015), we report NFI, GFI and AGFI further down. We relied on Pearson correlations to evaluate the GD's relationships to other instruments.

To check if it was adequate to aggregate the data at team level, we calculated $r_{wg(j)}$, ICC(1), and ICC(2), following LeBreton and Senter (2008). We required $r_{wg(j)}$ to be above .70, ICC(1) above .10, and ICC(2) above .30. According to LeBreton and Senter (2008), $r_{wg(j)} > .70$ indicates "strong agreement"; however, this interpretation depends on choosing an adequate distribution for the null hypothesis, as Biemann

et al. (2012) pointed out. We calculated $r_{wg(j)}$ based on an equal distribution as our null hypothesis, since none of the restrictions reported by Meyer et al. (2014) seemed applicable to our data.

Results

Confirmatory Factor Analysis

In the CFA, we accepted TLI at 0.95 or greater, RMSEA at 0.6 or lower (Hu and Bentler 1999), and χ^2/df ratio at 5 or lower (Schumacker and Lomax 2004). Thus, results confirmed the unidimensional structure of the GD questionnaire in both samples. NFI, GFI and AGFI were similar to those found by Navarro et al. (2015) in Brazil (Table 5).

Since the constructs selected for checking convergent validity showed high correlations among each other and with the GD (Table 3 and Table 4), we checked whether it was appropriate to measure these variables as separate constructs, or if all items rather represented one single common construct. A Common Latent Factor model did not converge due to a Haywood case. Thus, we conducted the following three additional CFAs with the members sample, including the items of the constructs GD, TCL mutual trust, TCL motivation, SABKWSE, and F-A-T solidarity to check the measurement model (MM):

1) The plain measurement model (MM 1 in Table 4) in which each item was only related to its respective construct.
2) A model (MM 2 in Table 4) which treated all available items of the constructs GD, TCL mutual trust, TCL motivation, SABKWSE, and F-A-T Solidarity as items of one single construct: general satisfaction with the team (GST).
3) For comparison, we ran MM 1 and added an exogenous variable (GST) that related to the endogenous constructs

GD, TCL mutual trust, TCL motivation, SABKWSE, and F-A-T Solidarity (CFA 3 in Table 5).

MM 1 (no common factor) showed the best fit, the model of MM 3 had a slightly worse fit, and the model of MM 2 (only one construct) showed much lower fit indices, even in the indices penalizing model complexity. Consequently, it was adequate to treat the measurements as separate constructs, even though it was likely to assume that a general factor had significant impact on all of them.

To evaluate the GD's convergent validity at item level, we calculated Composite Reliability (CR) and Average Variance Extracted (AVE). Convergent validity, acceptable with CR > .7 and AVE > .5 (Fornell and Larcker 1981), was partially confirmed, as AVE missed the quality criterion by .02, while CR met the defined quality standard (Table 6).

In both samples, Cronbach's α of the complete 8-item scale was greater than .80 (Table 6) and thus, the internal consistency was acceptable.

If the difference in CFI was .01 or smaller (Cheung and Rensvold 2002), we regarded models as invariant. In the sample of team members, the GD showed invariance between the group of researchers and the other job types, as well as between male and female participants: ΔCFI was .002 in both invariance tests. The leaders sample was too small for testing invariance. Although we found configural invariance between the Brazilian sample (Navarro et al. 2015) and our German sample, metric invariance was not confirmed with ΔCFI at .016.

Measurement Level

In the validation study by Navarro et al. (2016), all variables (including GD) were successfully aggregated at group level. We calculated $r_{wg(j)}$, ICC(1), and ICC(2) for all 133 teams in which we had measurements from at least two team members available. Mean team size was at $k = 3.06$. The results are shown in Table 7. For GD, $r_{wg(j)}$ was above .70 in all 133 teams. In 56 teams, we had a group score based on at least

Table 5 Model fit parameters

Country	Sample / Model	χ^2	df	χ^2/df	$p(\chi^2)$	TLI	RMSEA	NFI	GFI	AGFI
DE	Members CFA	35.32	18	1.96	.01	0.98	0.04	.98	.98	.97
	Leaders CFA	9.48	18	0.53	.95	1.06	0.00	.96	.98	.96
BR	Navarro et al. 2015	n.a.	n.a.	n.a.	n.a.	n.a.	0.01	.98	.99	.98
DE	Members MM 1	453.10	199	2.73	.00	0.95	0.06	.93	.91	.89
DE	Members MM 2	1338.94	209	6.41	.00	0.83	0.10	.83	.78	.73
DE	Members MM 3	589.28	204	2.89	.00	0.94	0.06	.92	.90	.88

Note. DE = Germany; BR = Brazil. χ^2 is the Chi-Square statistic (CMIN in Amos 22), and df is the respective number of degrees of freedom. $p(\chi^2)$ is the significance level of the χ^2 statistic, named P in Amos 22. TLI = Tucker-Lewis Index; RMSEA = Root Mean Square Error of Approximation; NFI = Normed Fit Index; GFI = Goodness-of-Fit Index; AGFI = Adjusted Goodness-of-Fit Index. n.a. = not available

Table 6 Data distribution, internal consistency and convergent validity

Mean$_1$	SD$_1$	α_1	α_2	CR$_1$	AVE$_1$
3.68	0.62	.87	.81	.88	.48

Notes. α is Cronbach's α. CR is Composite Reliability. AVE is Average Variance Extracted. Index 1 indicates data from the members' sample ($N = 501$), index 2 refers to data from the leaders' sample ($N = 104$)

two team members, and data submitted by the leader as well. In this sample, the Pearson correlation between the mean scores produced by the team members and the leader's score was $r = .25$ ($p = .07$). Apparently, the leaders had a positive bias on GD in their team: while the range of scores was [2.38, 4.50] for team members, it was [3.13, 5.00] for team leaders.

Convergent, Discriminant, and Concurrent Validity

As an additional evaluation of the GD scale's validity, we correlated the GD's sum-score to the scores resulting from the other instruments. We did this both at individual and at team level, since the aggregation of the data at team level had shown to be adequate, and since this allowed spotting possible inconsistencies across these levels.

The high correlation of the GD with the TCL dimension mutual trust demonstrates that it covers high-quality interpersonal relationships and proactive engagement towards team purposes. The GD's correlations with the SABKWSE items, reflecting group potency, and with the F-A-T factor solidarity were lower, yet relevant. The TCL factor democracy was the criterion variable that correlated lowest with the GD; democracy reflects a specific team process not directly covered by the GD instrument (Table 3). A very similar pattern was found when repeating the correlation analysis at team level (Table 4).

Concurrent validity of the GD was confirmed, showing high correlation ($p < .05$) with the criterion motivation and interest (Table 3, Table 4).

Discussion

The GD questionnaire is an attractive option for researchers who want to measure a groups' capacity to work together. It represents a psychosocial process that shows strong relationships with other theoretically related variables both in previous studies

(entitativity, group identification) and in this study (solidarity, mutual trust, motivation and interest in the group's tasks, group potency). It also relates to relevant outcome variables (team effectiveness, order and hygiene). The GD does not suffer from the limitations of other instruments, since it measures a continuous variable and not developmental group stages, and since it covers more key facets of well-developed groups than, for example, group cohesiveness. Results also show evidence for the intercultural applicability of the GD.

Main Findings

The results confirmed the construct validity of the German version of the GD. Its unidimensional structure was reproduced in the CFA. The GD scores among team members allowed for aggregating scores at group level. The GD also showed convergent and discriminant validity, correlating as supposed to similar groups constructs – both at an individual and a group level. The high correlation of the GD with the motivation of team members to engage in the team's tasks confirmed the concurrent validity. Despite the high correlations among other group constructs, additional CFAs confirmed the measurement model and showed that the items used represented distinct variables. The instrument showed measurement invariance regarding gender and job type. The good internal consistency is in line with results from other countries. Except for a marginally insufficient AVE, the instrument met all defined quality criteria.

The results show that the construct GD is furthermore applicable across different cultures. The characteristics of well-developed groups form a single measurement dimension in Spanish, Brazilian, and German samples. This indicates a broad applicability of the GD and may be relevant for developing theoretical models in teamwork investigation.

Since interrater agreement varied a lot between teams, we concluded that the GD is primarily an individual-level measurement, even though aggregation at group level is possible if teams with low $r_{wg(j)}$ scores are eliminated from the sample.

Limitations

We used two samples from only one research & development organization. To verify that our results are generalizable, we recommend gathering more data from diverse organizations and sectors in Germany. The overrepresentation of male

Table 7 Interrater agreement and Interrater reliability indices

	GD	TCL Mutual trust	TCL Democracy	FAT Solidarity	SABKWSE	TCL Mot. & Int.
$r_{wg(j)}$	.96	.75	.72	.86	.88	.90
ICC(1)	.14	.17	.18	.15	.11	.17
ICC(2)	.32	.38	.41	.35	.28	.39

Notes. $N = 133$ teams (408 participants); mean number of answered questionnaires per team $k = 3.06$

employees and of researchers was not critical, since the model was invariant across gender and job types. Nonetheless, other not-represented grouping variables could have a yet undetected influence on the measurement model. Furthermore, single-source bias may have inflated the correlation between the GD and its criterion variable of concurrent validity. Future research should include external data as validation criteria of successful teamwork, such as financial results or efficiency measures based on stakeholders' opinions. AVE was probably low because the scale reflects four semantically different aspects of well-developed groups. Although these aspects form a single factor, the items were not formulated to be parallel. In our view, this does not jeopardize the GD's validity, nor its applicability.

Implications for Future Research and Practice

The GD has a solid theoretical basis and makes reference to the relevant streams of research on group development and regarding the question what defines a team or workgroup. It continues the research line on groupality and takes aspects from the construct families of entitativity and groupness into account. It has also shown a stable factorial structure across different cultures and good reliability. We recommend using the GD to analyze the processes that make team members work together effectively. It can serve as a process variable in IPO models, or as a mediator in the more recently recommended IMOI (Input-Mediator-Output-Input) models. We recommend researching the factors that have a positive impact on GD, such as factors related to the organization, the task, the working environment, the leadership, etc. Regarding the GD itself, we recommend translating the instrument into more languages and to continue the validation, since this would increase its applicability in cross-cultural research.

The GD can help further research on the construct of group development, since it integrates the main shared properties of the three research branches on groupness, entitativity, and groupality in a single construct. In doing so, the GD does not mingle states and processes, as in the construct of entitativity (as defined by Lickel), to provide a defined and unidimensional assessment of group development as a process. Additionally, the GD overcomes the limitations of classic stage models of group development: in a sample as the one used in our study, which is mostly composed of teams that have a history of working together on projects in the past, models such as those presented by Tuckman and Jensen, or Wheelan and Hochberger, would be expected to differentiate rather poorly among such teams. Since the GD measures the extent to which the social element of a group of people working together has actually developed as a continuous measurement, we also gathered variance among not-recently formed teams. Moreover, previous research, cited above, has already shown that compared to Wheelan and Hochberger's (1996) GDQ, and that the GD explains additional variance of group effectiveness, and that the GD is a significant predictor for the

ultimate manager goals of group performance and group effectiveness. In summary, the GD is particularly helpful to predict group performance and to be used when most teams are expected to be already on a similar stage (e.g., performing).

To practitioners, we recommend using the GD for evaluating work teams, both because the instrument is quick and easy to use and because it meets high standards regarding validity and reliability. Being based on several key criteria of group functioning, it gives practitioners more clues on what to when low scores are measured, compared to stage-based instruments and measurements of group cohesion. The GD is a good predictor of team effectiveness and thus of high relevance to HR experts and management.

Summary

Summarizing, we recommend the German version of the GD questionnaire and propose its translation to more languages.

Data Policy

The datasets generated during and analyzed during the current study are available from the corresponding author on reasonable request.

Compliance with Ethical Standards

Conflict of Interest The authors declare that no conflict of interest exists; specifically, the authors do not benefit financially from the use other persons make of the tools validated in this study.

This study was carried out in accordance with the recommendations of the Ethical Principles of Psychologists and Code of Conduct (incl. The amendment from 2010) by the American Psychological Association (APA). Participation was voluntary and all participants gave written informed consent in accordance with the Declaration of Helsinki. The workers' committee of the organization checked our procedure for ethics and data privacy issues and approved the study.

References

Arrow, H., McGrath, J. E., & Berdahl, J. (2000). *Small groups as complex systems: Formation, coordination, development, and adaptation.* Thousand Oaks: SAGE Publications, Inc. https://doi.org/10.4135/9781452204666.

Biemann, T., Cole, M. S., & Voelpel, S. (2012). Within-group agreement: On the use (and misuse) of rWG and rWG(J) in leadership research and some best practice guidelines. *The Leadership Quarterly, 23*(1), 66–80. https://doi.org/10.1016/j.leaqua.2011.11.006.

Brodbeck, F., Guillaume, Y. R. F., & Winkler, M. (2010). Team climate for learning in higher education. WOP working paper no. 2010/1. http://www.psy.lmu.de/wirtschaftspsychologie/forschung/working_paper/wp_2010_1.pdf. Accessed 09 October 2019.

Carpenter, S., & Radhakrishnan, P. (2002). The relation between Allocentrism and perceptions of Ingroups. *Personality and Social Psychology Bulletin, 28*(11), 1528–1537. https://doi.org/10.1177/014616702237580.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal, 9*(2), 233–255. https://doi.org/10.1207/S15328007SEM0902_5.

Fornell, C., & Larcker, D. F. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of Marketing Research, 18*(1), 39. https://doi.org/10.2307/3151312.

Hackman, R. J. (1987). The design of work teams. In J. W. Lorsch (Ed.), *Handbook of organizational behavior* (pp. 315–342). Englewood Cliffs: Prentice-Hall.

Hamilton, D. L., & Sherman, S. J. (1996). Perceiving persons and groups. *Psychological Review, 103*(2), 336–355. https://doi.org/10.1037//0033-295X.103.2.336.

Hamilton, D. L., Sherman, S. J., & Lickel, B. (2013). Perceiving social groups: The importance of the Entitativity continuum. In C. Sedikides, J. Schopler, C. A. Insko, & C. Insko (Eds.), *Intergroup cognition and intergroup behavior* (pp. 56–83). New York: Psychology Press. https://doi.org/10.4324/9780203763575-9.

Hogg, M. A. (1993). Group cohesiveness: A critical review and some new directions. *European Review of Social Psychology, 4*(1), 85–111. https://doi.org/10.1080/14792779343000031.

Hogg, M. A., Turner, J. C., & Davidson, B. (1990). Polarized norms and social frames of reference: A test of the self-categorization theory of group polarization. *Basic and Applied Social Psychology, 11*, 77–100. https://doi.org/10.1207/s15324834basp1101_6.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal, 6*(1), 1–55. https://doi.org/10.1080/10705519909540118.

Ilgen, D. R., Hollenbeck, J. R., Johnson, M., & Jundt, D. (2005). Teams in organizations: From input-process-output models to IMOI models. *Annual Review of Psychology, 56*, 517–543. https://doi.org/10.1146/annurev.psych.56.091103.070250.

ITC. (2017). The ITC guidelines for translating and adapting tests (2nd edn): Version 2.4. https://www.intestcom.org/. Accessed 02.11.2018.

Kauffeld, S., & Frieling, E. (2001). Der Fragebogen zur Arbeit im Team (F-A-T). *Zeitschrift für Arbeits- und Organisationspsychologie A&O, 45*(1), 26–33. https://doi.org/10.1026//0932-4089.45.1.26.

Kenny, D. A. (2016). Measuring model fit. http://davidakenny.net/cm/fit.htm. Accessed 09 October 2019.

LeBreton, J. M., & Senter, J. L. (2008). Answers to 20 questions about interrater reliability and interrater agreement. *Organizational Research Methods, 11*(4), 815–852. https://doi.org/10.1177/1094428106296642.

Leuteritz, J.-P., Navarro, J., & Berger, R. (2017). How knowledge worker teams deal effectively with task uncertainty: The impact of transformational leadership and group development. *Frontiers in Psychology, 8.* https://doi.org/10.3389/fpsyg.2017.01339.

Lickel, B., Hamilton, D. L., Wieczorkowska, G., Lewis, A., Sherman, S. J., & Uhles, A. N. (2000). Varieties of groups and the perception of group entitativity. *Journal of Personality and Social Psychology, 78*(2), 223–246. https://doi.org/10.1037/0022-3514.78.2.223.

Marks, M. A., Mathieu, J. E., & Zaccaro, S. J. (2001). A temporally based framework and taxonomy of team processes. *The Academy of Management Review, 26*(3), 356. https://doi.org/10.2307/259182.

McGrath, J. E. (1984). *Groups, interaction and performance.* Englewood Cliffs: Prentice-Hall.

Meneses, R., Ortega, R., Navarro, J., & de Quijano, S. D. (2008). Criteria for assessing the level of group development (LGD) of work groups: Groupness, entitativity, and groupality as theoretical perspectives. *Small Group Research, 39*(4), 492–514. https://doi.org/10.1177/1046496408319787.

Meyer, R. D., Mumford, T. V., Burrus, C. J., Campion, M. A., & James, L. R. (2014). Selecting null distributions when calculating r wg: A tutorial and review. *Organizational Research Methods, 17*(3), 324–345. https://doi.org/10.1177/1094428114526927.

Miller, D. L. (2003). The stages of group development: A retrospective study of dynamic team processes. *Canadian Journal of Administrative Sciences / Revue Canadienne des Sciences de l'Administration, 20*(2), 121–134. https://doi.org/10.1111/j.1936-4490.2003.tb00698.x.

Moser, K. S., Schaffner, D., & Heinle, M. (2005). Entwicklung und Validierung einer bereichsspezifischen Skala zur Erfassung arbeitsbezogener kollektiver Wirksamkeitserwartungen [development and validation of an area-specific scale for the assessment of work-related collective expectations of efficacy]. *Zeitschrift für Arbeits- und Organisationspsychologie A&O, 49*(2), 85–91. https://doi.org/10.1026/0932-4089.49.2.85.

Navarro, J., Meneses, R., Miralles, C., Moreno, D. J., & Campelo, V. (2015). What distinguish teams from social aggregates?: A tool to assess the group development. *Anales de Psicología, 31*(3), 921–929. https://doi.org/10.6018/analesps.31.3.183831.

Navarro, J., Meneses, R., Nadal, M., & Landsberger, E. (2016). Desarrollo y desempeño en equipos de proyecto: validez incremental de la escala de desarrollo grupal [Development and performance in project teams: incremental validity of the group development scale]. *Anuario de Psicología, 46*, 8–16.

Roca Cortés, N. (2001). Grupos mediadores de la innovación profesional: estudio comparativo de procesos. *Revista de Psicología Social, 16*(2), 177–191. https://doi.org/10.1174/021347401317351125.

Schumacker, R. E., & Lomax, R. G. (2004). *A beginner's guide to structural equation modeling* (2nd ed.). Mahwah: Lawrence Erlbaum Associates.

Sharma, S., Mukherjee, S., Kumar, A., & Dillon, W. R. (2005). A simulation study to investigate the use of cutoff values for assessing model fit in covariance structure models. *Journal of Business Research, 58*(7), 935–943. https://doi.org/10.1016/j.jbusres.2003.10.007.

Tuckman, B. W., & Jensen, M. A. C. (1977). Stages of small-group development revisited. *Group & Organization Management, 2*(4), 419–427. https://doi.org/10.1177/105960117700200404.

Tyupa, S. (2011). A theoretical framework for back-translation as a quality assessment tool. New voices in Translation Studies. (7), 35–46. Retrieved from https://core.ac.uk/download/pdf/53121361.pdf

Wheelan, S. A., & Hochberger, J. M. (1996). Validation studies of the group development questionnaire. *Small Group Research, 27*(1), 143–170. https://doi.org/10.1177/1046496496271007.