SM 1f: Details and examples of moderator coding

Moderator	**Notes from coding instructions** *(with examples where helpful)*
Diversity dimension	• *Nationality* coded rather than *ethnicity* where both were possible (e.g., % Chinese) *Specific sub-dimensions could be further disambiguated in future research:* • *Functional* diversity may both reflect current function and previous function (e.g., in top management teams) • *Tenure* diversity may concern tenure in the specific team or in the organization.
Time	Year of data collection (if reported), otherwise year of publication. • If diversity and performance measures were collected in different years, this concerned performance measures. • If data was collected over multiple years, the midpoint of the range was used here.
Task complexity	*We operationalized this using required education to have an objective criterion and attain adequate interrater reliability. Nevertheless, it still refers to the task, not the individuals; particularly highly educated individuals can still engage in low complexity tasks.* High (operationalized as requiring graduate degree, or similarly intensive training), e.g., elite sports (Amenta et al., 2012), academic research (Hinnant et al., 2012), top management (Hambrick et al., 2014) Medium (operationalized as requiring college degree), e.g., nursing teams (Chang et al., 2006), complex technical manufacturing (Kirkman & Shapiro, 2005), undergraduate student projects (To et al., 2021) Low (operationalized as not requiring college degree), e.g., standardized manufacturing (e.g., garments, Boehm et al., 2021), fast food restaurant customer service (Sacco, 2002) or simple experimental tasks (Przybysz et al., 2014)
Interdependence	Low (pooled or sequential work), e.g., retail sales (Ely, 2004), "brainwriting" (Michinov & Jeanson, 2021) Medium (reciprocal turn-taking), e.g., Wikipedia editing (Ren et al., 2016),baseball teams (Bezrukova et al., 2016) High (more intense collaboration), e.g., top management teams (del Carmen Triana et al., 2019), soccer teams (Nüesch, 2009)
Country	Country where performance was achieved. • Where not reported but all authors came from same country, coded that country as location.
Success criterion	Divergence (e.g., many ideas), for instance Convergence (e.g., best idea), for instance Production (of pre-defined product) – for instance, time taken for garment production () *Note that this could not be coded for financial measures, or for creativity measures mixing the number and quality of ideas.*
Diversity climate	Positive (> midpoint on measurement scale, e.g., Ali & Konrad, 2017, *or* experimentally induced, e.g., Joshi, 2002) Negative (< midpoint on measurement scale *or* experimentally induced) – *never* observed Not reported – *most common*
Psychological safety	High (> midpoint on measurement scale, e.g., Liu et al., 2021, *or* experimentally induced – *never observed*) Low (< midpoint on measurement scale *or* experimentally induced) – *never* observed Not reported – *most common*

Authority differentiation	High (i.e. one/few team members take decisions), e.g., SWAT teams under a commander (Putney, 2003) Mixed (e.g., team leaders appointed by the team, and roles agreed), e.g., in a student start-up (Hoogendoorn et al., 2012) Low (i.e. consensual approach to decision-making), e.g., business plan competition teams without set roles (Foo et al., 2005) *Mostly codable for students engaging in business simulations, otherwise rarely described.*
Team longevity	Unit best describing lifespan of team until performance was measured: hours/days/weeks/months or years (defined as unit in which typical lifespan is at least 2, i.e. 6 weeks coded as weeks, 9 weeks coded as months). Ongoing teams (e.g., top-management teams of established companies) coded as stable.
Team virtuality	Very few explicitly virtual teams, though there was often no information on co-location Physical co-presence with face-to-face interaction mostly coded for experiments or sports teams
Diversity measure	*Coded based on diversity measure used:* Variety (e.g., Blau index, Teachman index, count of categories) – maximum when each team member has distinct identity Separation (e.g., coefficient of variation, SD, mean Euclidian distance) – maximum when team polarized into subgroups Other (e.g., binary splits into diverse and homogenous teams) *NB: Choice between variety and separation of little import for teams consisting of two categories – so future work may want to combine this with the range / number of categories.*
Performance measure	Objective – includes achievement of quantitative targets (e.g., sales or production quotas) Subjective (by team members) Subjective (by team leader) Subjective (by external rater) – includes awards and expert judgements
Study design	Observational: most studies, frequently either survey-based or based on secondary data (e.g., financial databases) Experimental: largely student business simulations, e.g., Aggarwal et al. (2023) Quasi-experimental (**dropped**, as this is about analysis more than design)
Article focus	Is link between diversity and performance: focal hypothesis / auxiliary hypothesis / descriptive result *Coding rules:* • Focal hypotheses will usually appear in title or at least abstract • Auxiliary hypotheses would only appear in long list of hypotheses not all concerning diversity • Descriptive results may stem from covariates, or when diversity is used as a moderator • When hypotheses are only about moderated/conditional effects of diversity, the correlation is coded as auxiliary.
Citation count	Retrieved from Google Scholar (as no other sources covers all included languages), reports not found in Google Scholar counted as having 0 citations