Using rating scales, LLMs, and frameworks to generate items for a FACETS construct study

Arno Klein (2-8 Feb 2026)

This document describes the preparation for a study designed to understand which factors teachers consider to be the most important aspects of their students' development and functioning. Their responses will help inform the development of educational assessment tools. They will be presented with pairs of student qualities (such as "Self-Control" vs. "Empathy") and asked to choose which of the two they consider to be a more relevant and important characteristic of students with whom they work.

1. Manually extract thousands of questions from 153 rating scales

I wanted to include candidate constructs from the 153 assessments used in the HBN and NKI-RS1/2 studies. I manually reduced the number of items down from 5,632 to 2,936 by removing as many redundant and "other" questions as I could and removed or reduced long checklists dealing with:

1. Drugs
2. Activities or foods (removed many entries in long checklists
3. Scary or traumatic events observed or experienced
4. Risk-taking behaviors
5. Which hand is used
6. Muscle cramps
7. Suicidal ideation/attempts
8. Thoughts about weight
9. Motor/vocal tics
10. Demographics and heritage culture

2. Generate items from rating scales

An LLM generated "FACETS-appropriate" items from the rating scale questions (see prompt).

Claude Opus 4.5 LLM prompt (run 3 times)

```
I am creating a non-clinical assessment of children's emotional, behavioral, and cognitive health and well-being from the perspective of, and for the use of, a teacher. Use information from the assessment/questions csv file (containing thousands of items from many clinical rating scales) to generate a set of categories, subcategories, brief descriptions, formatted as follows: Source, Category, Subcategory, Brief Description, Description
```

- The resulting assessment must have between 5 and 20 categories and much fewer than 100 items. Do not try to artificially balance or constrain the number of subcategories within a category.
- Generate nonclinical sub/categories pertaining to a child's emotional, behavioral, and cognitive health and well-being. Avoid clinical terms and medical conditions. Do not attempt to align sub/categories with any particular theoretical framework or diagnostic system (e.g., DSM-5, IDEA disability categories, CASEL SEL competencies, RDoC).
- Focus on domains most relevant to students, whether observable or not (e.g., attention, behavior, emotional regulation, social functioning, academics, etc.). That is, include constructs that may require inference about internal states (e.g., self-esteem, anxiety symptoms).
- Do not include fine-grained subcategories. For example, "attention" is reasonable, but not subcategories like "sustained attention," "selective attention," "divided attention," etc.
- The categories should be developmentally universal (K-12).

Input Step 1: 2,936 questions

Output 105 items generated by three LLM prompt submissions:

- 13 categories and 43 items
- 10 categories and 28 items
- 12 categories and 34 items

3. Generate items using an LLM

An LLM generated “FACETS-appropriate” items from scratch (see prompt).

Claude Opus 4.5 LLM prompt (run 3 times)

```
I am creating a non-clinical assessment of children's emotional, behavioral, and cognitive health and well-being from the perspective of, and for the use of, a teacher. Create a csv file with columns: Source (name of LLM), Category, Subcategory, Brief Description, Description
```

- The resulting assessment must have between 5 and 20 categories and much fewer than 100 items. Do not try to artificially balance or constrain the number of subcategories within a category.
- Generate nonclinical sub/categories pertaining to a child's emotional, behavioral, and cognitive health and well-being. Avoid clinical terms and medical conditions. Do not attempt to align sub/categories with any particular theoretical framework or diagnostic system (e.g., DSM-5, IDEA disability categories, CASEL SEL competencies, RDoC).
- Focus on domains most relevant to students, whether observable or not (e.g., attention, behavior, emotional regulation, social functioning, academics, etc.). That is, include constructs that may require inference about internal states (e.g., self-esteem, anxiety symptoms).
- Do not include fine-grained subcategories. For example, “attention" is reasonable, but not subcategories like “sustained attention,” “selective attention,” "divided attention,” etc.
- The categories should be developmentally universal (K-12).

Output Step 2: 124 items generated by three LLM prompt submissions:

- 12 categories and 38 items
- 12 categories and 41 items
- 12 categories and 45 items

4. Extract all items from 40 frameworks

An LLM scraped the http://exploresel.gse.harvard.edu/frameworks/ website to extract all of the Terms and Definitions from all 40 SEL frameworks. I removed the less relevant EDC Work Ready Now! Framework.

LLM prompt

```
http://exploresel.gse.harvard.edu/frameworks/ contains frameworks. Each framework contains terms and definitions. For example, the "21st Century Learning" framework (http://exploresel.gse.harvard.edu/frameworks/3/terms/) contains the term "think creatively" with the definition "Use a wide range of idea creation techniques (such as brainstorming); Create new and worthwhile ideas (both incremental and radical concepts); Elaborate, refine, analyze and evaluate their own ideas in order to improve and maximize creative efforts". Create a csv file with columns "Framework", "Term", and "Definition" for all terms for all the frameworks.
```

Output 523 terms and definitions in 40 frameworks extracted from the Harvard website

5. Manually aggregate items from all sources

Output 876 items from FACETS, rating scales, LLMs, and frameworks

6. Manually compare items from all sources

I went carefully through all 876 items from all four sources (FACETS, frameworks, rating scales, and LLMs), and for each non-FACETS item, manually determined whether it was included in FACETS, coarser category of a FACETS item, more fine-grained category of a FACETS item, not included in FACETS, or irrelevant. I did not rely on the similarity measures calculated below — I used them just to help me sort items to make it easier to compare non-FACETS items with FACETS items. Similarity measures were based on sentence-BERT embeddings (Reimers & Gurevych, 2019), specifically the all-mpnet-base-v2 model pre-trained on over 1 billion sentence pairs. For each item, we concatenated the name, brief description, and full description into a unified text representation and computed cosine similarity between embedding vectors.

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-Networks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) (pp. 3982-3992).

Claude Opus 4.5 LLM prompt

I am creating a non-clinical assessment of children's emotional, behavioral, and cognitive health and well-being from the perspective of, and for the use of, a teacher. See the csv file for current items (Subcategories). Use analyze_similarities_sbert.py to find similarities (0-1 scores) and redundancies across all subcategories and generate a duplicate csv file with three new columns:

(1) “Max FACETS similarity value” column with a maximum similarity measure between row and any FACETS row.

(2) “Max FACETS similarity” column with the Subcategory with the maximum similarity measure between row and any FACETS row.

(3) “Max non-FACETS similarity value” column with a maximum similarity measure between row and any other non-FACETS row.

(4) “Max non-FACETS similarity” column with the Subcategory with the maximum similarity measure between row and any other non-FACETS row.

Make sure to follow the guidelines below:

- The subcategories should be nonclinical and should pertain to a child's emotional, behavioral, and cognitive health and well-being (they avoid clinical terms and medical conditions). These subcategories are not intended to align with any particular theoretical framework or diagnostic system (e.g., DSM-5, IDEA disability categories, CASEL SEL competencies, RDoC).
- The subcategories should focus on domains most relevant to students, whether observable or not (e.g., attention, behavior, emotional regulation, social functioning, academics, etc.). That is, they include constructs that may require inference about internal states (e.g., self-esteem, anxiety symptoms).
- Don’t include fine-grained subcategories. For example, “attention" is reasonable, but not subcategories like “sustained attention,” “selective attention,” "divided attention,” etc.
- The subcategories should be developmentally universal (K-12).

Input Step 5: 876 items from FACETS, rating scales, LLMs, and frameworks

Output 876 items with similarity measures

Flag	Count	Meaning	Implication
[no entries]		FACETS items	
irrelevant = 1	73	Item should be excluded	Exclude
misattributed FACETS = 1	72	"Max FACETS similarity" is WRONG	Can't trust direct assignment
not in FACETS = 1	323	No good FACETS equivalent exists	May not belong to any anchor
coarse FACETS item = 1	97	Broader than its FACETS match	Valid, include
granular FACETS item = 1	298	More specific than its FACETS match	Valid, include

7. Manually consolidate relevant, non-redundant items

I manually consolidated all relevant non-FACETS items not included in FACETS, removing redundancies.

Input Step 6: 876 items with similarity measures

Output 146 consolidated items

8. Assign all 803 relevant items to 146 consolidated items

An LLM assigned each of the 803 relevant items to the 146 consolidated items.
Assignment procedure (see “facets - 8. LLM_assignment_method.md”):

> For each candidate item:
>
> 1. **Construct prompt** containing:
> - The item's name, brief description, and full description
> - Complete list of 146 FACETS anchors with brief descriptions
>
> 2. **Query LLM** with instruction:
> > "Which FACETS category does this item best match? The item should measure the same or a very closely related construct."
>
> 3. **Parse response** containing:
> - `assigned_anchor`: Exact name of matched FACETS category, or null if no good match
> - `confidence`: "high" (same construct), "medium" (closely related), or "low" (weak match)
> - `reasoning`: One-sentence explanation
>
> 4. **Validate anchor name** against the list of valid FACETS names

Claude Opus 4.5-generated LLM prompt

```
I have psychological/educational assessment constructs that need to be assigned to standardized categories.

**Input files:**
1. "facets - 7. consolidated (146).csv" - 146 FACETS anchor categories with columns: Subcategory, Source, Category, Brief Description, Description
2. "facets - 6. similarities (876).csv" - 876 candidate items to assign, with columns including: Subcategory, Source, Category, Brief Description, Description, irrelevant (1 = exclude)

**Task:**
Create a Python script that uses an LLM (Claude) to assign each candidate item to one of the 146 FACETS anchors.

**Method:**
1. Filter out items where irrelevant = 1
2. For EACH of the ~803 remaining items, make ONE LLM API call:
  - Show the item's name and description
  - Show the list of 146 FACETS anchors with their brief descriptions
  - Ask: "Which FACETS category does this item best match?"
  - Get response with: assigned_anchor, confidence (high/medium/low), reasoning

3. Group results by assigned anchor and output CSV with columns:
  - Items: comma-delimited list (duplicates as "Name (count)")
  - Sources: comma-delimited as "[Source] ([Category])"
  - Brief Descriptions: "///"-delimited
  - Descriptions: "///"-delimited

**Requirements:**
- One LLM call per item (not batched) for full attention to each assignment
- Save progress incrementally (can resume if interrupted)
- Track confidence levels for quality assurance
- Handle items with no good match (assign to null)
- Output both the grouped CSV and a JSON file with all individual assignments

**Why one-at-a-time:**
- Each item gets full context and attention
- More accurate than batching (no confusion between items)
- Auditable individual decisions
- Cost is minimal (~$3-5 for 800 calls)

Please also provide:
1. A publication-ready method description
2. The prompt that could recreate this in a new session
```

Input Step 6: 876 items with similarity measures and Step 7 146 consolidated items

Output 146 item groups

Metric	Value
Total assignments	657
High confidence	624 (95.0%)
Medium confidence	30 (4.6%)
Low confidence	3
No match	4*
Items per anchor	Min: 1, Max: 23, Mean: 4.2

*takes responsibility for professional growth, trust, roles, barriers

Example: Time management

- Time Management
- Time Awareness
- manages time
- Managing Deadlines

Note: "Screen Time" stays separate because it measures device usage, not time management skills - a more accurate semantic distinction than the SBERT similarity approach.

9. Manually select and rename items

Bennett and I manually chose distinct and relevant items within each item group, renaming for clarity and consistency.

Input Step 8: 146 item groups

Output 1 93 items from 146 item groups

10. Candidate items

Bennett and I manually assigned 2 synonyms per item, with help from an LLM.

Claude Opus 4.5 LLM prompt to seed synonyms

```
Generate three synonyms (columns Synonym 1-3) for each Item in “facets - 10. candidate items (104).csv”, none of which share any words with any of the other two synonyms or with the original Subcategory, and output a new csv file.
```

Input Step 9: 93 items from 146 item groups

Output 96 candidate items plus synonyms

Study: Pairwise importance judgments for construct validation

Research Questions

1. Importance hierarchy: Which constructs do teachers prioritize when reflecting on their students?
2. Construct dependence: Which constructs are perceived as interchangeable vs. distinct?
3. Construct validity (with synonyms): Do target items and their synonyms show equivalent importance patterns?

Methods

Participants: K-12 teachers with English as their primary language

Stimuli: 90 target items (candidate constructs), 2 synonyms per item

Task per trial:

1. Present two items (e.g., "Self-Control" vs. "Empathy")
2. Prompt: *"When considering your students,*
   which of the two presented terms
   is more relevant and important
   for you to understand or assess?
   Please use your professional judgment."
3. Forced binary choice between item x item or item x synonym (but not synonym x synonym)

Design

If we randomly assign 26 constructs per participant:

- 26-choose-2 = 325 cross-construct pairs (complete within-subset design)
- 1 randomly selected synonym per construct = 26 within-construct pairs
- Each participant completes 325 + 26 = 351 comparisons

At **5s per pair**: 351/12 ≈ **29 minutes per participant**.
For every 5 minutes, 1 attention check (nonsense comparison), adding 30s.
Each of **~300 participants** receives 26 of 90 constructs without replacement.

- Each construct appears in ~300 × 26/90 ≈ 87 participants' subsets.
- Each cross-construct pair observed by ~300 × C(26,2)/C(90,2) ≈ 24 independent judges.
- Each synonym tested ~44 times (both synonyms covered across participants).

24 obs/pair: sufficient for stable Bradley-Terry estimation (≥0.5 logit differences).
87 obs/construct: sufficient to test synonym equivalence (P = 0.50) to ±8 pp at α = .05.
Complete within-subset design avoids cross-set linking problems and need for hierarchical priors.
$12/hr × ½ hr × 300 participants = $1,800 + Prolific fees.

Analysis Plan

1. Importance ranking (Bradley-Terry Model):
   The Bradley-Terry model estimates latent "worth" parameters for each item from pairwise choices. The output is a ranked list of 90 constructs by teacher-perceived importance, with confidence intervals.
2. Construct dependence (choice consistency analysis):
   Within-pair variability: If teachers are split 50/50 on a pair, the items may be equally important or conceptually overlapping (hard to distinguish).
   - Transitivity violations: If A > B and B > C but C > A, this circular pattern suggests the three items may be measuring related constructs.
   - Co-win/co-loss patterns: Items that consistently beat (or lose to) the same opponents likely measure similar constructs.
   - Analytical approach: Cluster items by residual covariance patterns from Bradley-Terry predictions.
3. Construct validity via synonyms
   Hypothesis: A target item and its synonym should have:
   - Similar Bradley-Terry scores (equal importance)
   - Highly variable direct comparison (50/50 when pitted against each other)
   - Similar patterns of wins/losses against other items

   Discriminant validity: Target items should clearly beat/lose to *non*-synonyms with higher consistency than they beat/lose to synonyms.

   Metric: Compute correlation of win-rates between target and synonym.

   A high correlation means good construct validity.

Expected Outputs

1. Importance ranking of 90 constructs (with CIs)
2. Dependence matrix: which constructs are perceived as interchangeable
3. Validity coefficients: for each target-synonym pair
4. Dimensionality estimate: how many latent factors underlie teacher priorities
5. Problematic items: those with high inconsistency or poor synonym agreement