Basado en la literatura del corpus, específicamente en los trabajos de **Jang (2015)**, **Börner et al. (2018)** y **Garcia et al. (2024)**, la respuesta es afirmativa: **es metodológicamente válido y altamente recomendado** agrupar competencias específicas en macro-categorías semánticas o dimensiones latentes para tratar la dispersión (*sparsity*) y la alta dimensionalidad, especialmente para mejorar la interpretabilidad y la señal en los modelos. A continuación, se detalla la postura de estos autores sobre el tratamiento de vectores de habilidades:

1. Hyewon Jang: Agrupación Semántica sobre Reducción Estadística

Jang aborda directamente el problema de la dimensionalidad al tratar con 109 descriptores de habilidades, conocimientos y actividades laborales extraídos de la base de datos O*NET.

- **Problema con métodos puramente estadísticos:** Jang intentó inicialmente utilizar el **Análisis de Componentes Principales (PCA)** para condensar los datos. Sin embargo, encontró que el PCA era problemático porque requería demasiados componentes para explicar la varianza y, crucialmente, los componentes resultantes estaban "desalineados" (*mismatched*) con interpretaciones significativas (por ejemplo, mezclando habilidades físicas con otras que no tenían relación lógica para el perfil STEM) 1, 2.
- **Solución recomendada (Agrupación Semántica):** Para resolver esto, Jang recomendó y aplicó una **categorización semántica basada en marcos teóricos**. Utilizó el marco de Katz y Kahn (1978) para agrupar 52 descriptores específicos (como "Pensamiento Crítico" o "Programación") en **5 macro-categorías** o competencias clave (ej. "Habilidades de resolución de problemas mal definidos", "Habilidades de comunicación social", "Habilidades técnicas y de ingeniería") 3, 4, 5.
- **Validación:** Para asegurar que esta agrupación fuera válida y no arbitraria, utilizó la verificación de la **fiabilidad inter-evaluador** (*inter-rater reliability*), logrando un coeficiente Kappa de Cohen de 0.74, lo que valida metodológicamente la agrupación manual basada en definiciones operativas 6.

2. Katy Börner et al.: Uso de Taxonomías Jerárquicas

Börner se enfrentó a un problema de dispersión masiva al analizar **13,218 habilidades únicas** extraídas de millones de ofertas de trabajo y currículos.

- **Estrategia de Taxonomía:** Para manejar esta hiper-dimensionalidad, los autores no utilizaron las habilidades como vectores aislados, sino que las anclaron a la **taxonomía de habilidades de Burning Glass (BG)**. Esta taxonomía organiza las 13,000+ habilidades en **560 clústeres de habilidades**, que a su vez se agregan en **28 familias de habilidades** 7.
- **Recodificación para Análisis:** Börner validó la metodología de recodificar estas habilidades específicas en macro-categorías binarias más amplias, distinguiendo explícitamente entre habilidades **"Hard"** (cuantitativas/técnicas) y **"Soft"** (sociales/comunicación) para facilitar la visualización y el análisis de redes de co-ocurrencia 8. Esto demuestra que agrupar competencias en familias semánticas es esencial para mapear la oferta y la demanda en "espacios de habilidades" complejos.

3. Garcia et al. y otros: Reducción Dimensional Matemática (PCA)

A diferencia de Jang, otros autores en el corpus defienden la validez de la agrupación estadística cuando el objetivo es puramente predictivo o cuando se busca simplificar el análisis sin perder información estadística.

- **Garcia et al. (2024):** Utilizaron **Análisis de Componentes Principales (PCA)** para condensar múltiples variables de resultados y habilidades en dos componentes principales: "Habilidades Personales" y "Habilidades Profesionales". Argumentan que el PCA es excelente para condensar información de muchas variables interrelacionadas y simplificar el análisis sin una pérdida significativa de información, revelando estructuras latentes 9, 10.
- **Zhao et al. (2022):** En el contexto de la predicción de movilidad laboral, utilizaron PCA para reducir la dimensionalidad de los datos de entrada de **22 dimensiones a 9 dimensiones**. Argumentan que un número excesivo de entradas irrelevantes (dispersión/ruido) puede resultar fácilmente en el sobreajuste (*overfitting*) del modelo durante el entrenamiento. La reducción a componentes principales retuvo la información esencial y mejoró la eficiencia experimental 11, 12.

Conclusión y Recomendación para su Tesis

Basado en el corpus, la respuesta a su pregunta es:

- **Sí, es metodológicamente válido** agrupar competencias en macro-categorías. De hecho, es necesario para evitar el *overfitting* y la maldición de la dimensionalidad en datasets pequeños o dispersos.
- **Estrategia Recomendada:**
- Si su prioridad es la **interpretabilidad educativa** (saber exactamente qué enseñar), siga el enfoque de **Jang**: agrupe las habilidades manualmente en categorías teóricas (ej. "Competencias Técnicas", "Habilidades Blandas") y valide la agrupación mediante revisión de expertos o fiabilidad inter-evaluador.
- Si su prioridad es maximizar la **potencia predictiva** estadística y descubrir patrones latentes no obvios, siga el enfoque de **Garcia** o **Zhao**: utilice PCA para reducir el vector de habilidades a sus componentes principales.

Para su caso específico (datasets pequeños), la **agrupación semántica (tipo Jang)** suele ser más robusta al ruido que el PCA, ya que el PCA requiere un tamaño de muestra considerable para que las correlaciones sean estables. Agrupar, por ejemplo, "Python", "Java" y "C++" bajo la etiqueta "Programación" reducirá la dispersión (muchos ceros en el vector) y aumentará la señal de la variable para el modelo.